A New Algorithm: Efficient Cost Algorithms For Reducing High Toll Transactions

Vinay Singh[1] Dr. S.Gavaskar[2]
[1,2] Department of Computer Science, Galgotias University
Greater Noida, India

Abstract---Mining Cost Efficient item-sets from a transactional database refers to the discovery of transaction sets with Cost Efficient characteristics improving all profits. Although a number of important algorithms have been proposed in recent years, they incur the problem of producing a large number of candidate item-sets for Cost Efficient item-sets. Such a large number of candidate item-sets degrade the mining performance in terms of execution time and space requirement. The condition may become worse when the database contains lots of long transactions or long Cost Efficient item-sets. In this paper, we propose new algorithm, for mining Cost Efficient item-sets with a set of effective strategies for pruning candidate item-sets. The information of Cost Efficient item-sets is maintained in a tree-based data structure such that candidate item-sets can be generated efficiently with only two scans of database.

Keywords:-Data Mining, High utility item sets Transactional databases.

I. INTRODUCTION

Data mining is an interdisciplinary field, the confluence of a set of disciplines, including database systems, statistics, machine learning, visualization, and information science (Figure Moreover, depending on the data mining approach used, techniques from other disciplines may be applied, such as neural networks, fuzzy and/or rough set theory, knowledge representation, inductive logic programming, or high-performance computing. Depending on the kinds of data to be mined or on the given data mining application, the data mining system may also integrate techniques from spatial data analysis, information retrieval, pattern recognition, image analysis, signal processing, computer graphics, Web technology, economics, business, bioinformatics, or psychology. Because of the diversity of disciplines contributing to datamining, data mining research is expected to generate a large variety of data mining systems. Therefore, it is necessary to provide a clear classification of data mining systems, which may help potential users distinguish between such systems and identify those that best match their needs. Data mining systems can be categorized according to various criteria, as follows:

A. Clustering :

Clustering is the task of partitioning the points into natural groups called clusters, such that points within a group are very similar, whereas points across clusters are as dissimilar hierarchical, density-based, graph-based and spectral clustering. It starts with representative-based clustering methods which include the K-means and Expectation-Maximization (EM) algorithms. K-means is a greedy algorithm that minimizes the squared error of points from their respective cluster means, and it performs hard clustering, that is, each point is assigned to only one cluster. Classification The classification task is to predict the label or class for a given unlabeled point. Formally, a classifier is a model or function M that predicts the class label ˆy for a given input example x, that is, ˆy = M(x), where ˆy ? {c1, c2, · · · ,ck} is the predicted class label (a categorical attribute value). To build the model we require a set of points with their correct class labels, which is called a training set.

B. Classification:

The classification task is to predict the label or class for a given unlabeled point. Formally, a classifier is a model function M that predicts the class label ˆy for a given input example x, that is, ˆy = M(x), where ˆy ? {c1, c2, · · · ,ck} is the predicted class label (a categorical attribute value). to be estimated which scales as O(d2). The naive Bayes classifier makes the simplifying assumption that all attributes are independent, which requires the estimation of only O(d we require a set of points with their correct class labels, which is called a training set.the join) parameters. It is, however, surprisingly effective for many datasets. In Chapter 19 we consider the popular decision tree classifier, one of whose strengths is that it yields models that are easier to understand compared to other methods.

The support vector machine (SVM) is one of the most effective classifiers for many different problem domains. The goal of SVMs is to find the optimal hyperplane that maximizes the margin between the classes. Via the kernel trick SVMs can be used to find non-linear boundaries, which nevertheless correspond to some linear hyperplane in some high-dimensional "non-linear" space. One of the important tasks in classification is to assess how good the models are.

C. Transactional Data Mining :

Association rule mining (ARM) identifies frequent itemsets from databases and generates association rules by considering each item in equal value. However, items are actually different in many aspects in a number of real applications, such as retail marketing, network log, etc. The difference between items makes a strong impact on the decision making in these applications. Therefore, traditional ARM cannot meet the demands arising from these applications. By considering the different values of individual items as utilities, utility mining focuses on identifying the itemsets with high utilities. As "downward closure property" doesn't apply to utility mining, the generation of candidate itemsets is the most costly in terms of time and memory space.

D. Issues in Transaction Mining:

The scope of this book addresses major issues in data mining regarding mining methodology, user interaction, performance, and diverse data types. These issues are introduced below:

E. Mining Methodology And User Interaction Issues:
These reflect the kinds of knowledge mined, the ability to mine knowledge at multiple granularities, the use of domain knowledge, ad hoc mining, and knowledge visualization. Mining different kinds of knowledge in databases: Because different users can be interested in different kinds of knowledge, data mining should cover a wide spectrum of data analysis and knowledge discovery tasks, including data characterization, discrimination, association and correlation analysis, classification, prediction, clustering, outlier analysis, and evolution analysis (which includes trend and similarity analysis).

F. Incorporation Of Background Knowledge:
Background knowledge, or information regarding the domain under study, may be used to guide the discovery process and allow discovered patterns to be expressed in concise terms and at different levels of abstraction. Domain knowledge related to databases, such as integrity constraints and deduction rules, can help focus and speed up a data mining process, or judge the interestingness of discovered patterns.

G. Data Mining Query Languages And Ad Hoc Data Mining:
Relational query languages (such as SQL) allow users to pose ad hoc queries for data retrieval. In a similar vein, high-level data mining query languages need to be developed to allow users to describe ad hoc data mining tasks by facilitating the specification of the relevant sets of data for analysis, the domain knowledge, the kinds of knowledge to be mined, and the conditions and constraints to be enforced on the discovered patterns. Such a language should be integrated with a database or data warehouse query language and optimized for efficient and flexible data mining.

H. Presentation And Visualization Of Data Mining Results:
Discovered knowledge should be expressed in high-level languages, visual representations, or other expressive forms so that the knowledge can be easily understood and directly usable by humans. This is especially crucial if the data mining system is to be interactive. This requires the system to adopt expressive knowledge representation techniques, such as trees, tables, rules, graphs, charts, crosstabs, matrices, or curves.

I. Handling Noisy Or Incomplete Data:
The data stored in a database may reflect noise, exceptional cases, or incomplete data objects. When mining data regularities, these objects may confuse the process, causing the knowledge model constructed to over fit the data. As a result, the accuracy of the discovered patterns can be poor. Data cleaning methods and data analysis methods that can handle noise are required, as well as outlier mining methods for the discovery and analysis of exceptional cases.

J. Pattern Evaluation—The Interestingness Problem:
A data mining system can uncover thousands of patterns. Many of the patterns discovered may be uninteresting to the given user, either because they represent common knowledge or lack novelty. Several challenges remain regarding the development of techniques to assess the interestingness.

K. Performance Issues:
1) Efficiency And Scalability Of Data Mining Algorithms: To effectively extract information from a huge amount of data in databases, data mining algorithms must be efficient and scalable. In other words, the running time of a data mining algorithm must be predictable and acceptable in large databases. From a database perspective on knowledge discovery, efficiency and scalability are key issues in the implementation of data mining systems. Many of the issues discussed above under mining methodology and user interaction must also consider efficiency and scalability.

2) Parallel, Distributed, And Incremental Mining Algorithms: The huge size of many databases, the wide distribution of data, and the computational complexity of some data mining methods are factors motivating the development of parallel and distributed data mining algorithms. Such algorithms divide the data into partitions, which are processed in parallel. The results from the partitions are then merged.Moreover, the high cost of some data mining processes promotes the need for incremental data mining algorithms that incorporate database updates without having to mine the entire data again "fromscratch." Such algorithms perform knowledge modification incrementally to amend and strengthen what was previously discovered.

Issues Relating To The Diversity Of Database Types: Handling of relational and complex types of data:
Because relational databases and data warehouses are widely used, the development of efficient and effective data mining systems for such data is important. However, other databases may contain complex data objects, hypertext and multimedia data, spatial data, temporal data, or transaction data. It is unrealistic to expect one system to mine all kinds of data, given the diversity of data types and different goals of data mining. Specific data mining systems should be constructed for mining specific kinds of data. Therefore, one may expect to have different data mining systems for different kinds of data.

3) *Mining Information From Heterogeneous Databases And Global Information Systems*: Local- and wide-area computer networks (such as the Internet) connect many sources of data, forming huge, distributed, and heterogeneous databases. The discovery of knowledge from different sources of structured, semi-structured, or unstructured data with diverse data semantics poses great challenges to data mining. Data mining may help disclose high-level data regularities in multiple heterogeneous databases that are unlikely to be discovered by simple query systems and may improve information exchange and interoperability in heterogeneous databases. Web mining, which uncovers interesting knowledge about Web contents, Web structures, Web usage, and Web dynamics, becomes a very challenging and fast-evolving field in data mining.

II. RELATED WORK

Liu, Mengchi, et al[1]. High utility itemsets refer to the sets of items with high utility like profit in a database, and efficient mining of high utility itemsets plays a crucial role in many reallife applications and is an important research issue in data mining area. To identify high utility itemsets, most existing algorithms first generate candidate itemsets by overestimating their utilities, and subsequently compute the exact utilities of these candidates. These algorithms incur the problem that a very large number of candidates are generated, but most of the candidates are found out to be not

high utility after their exact utilities are computed. In this paper, we propose an algorithm, called HUI-Miner (High Utility Itemset Miner), for high utility itemset mining. HUI-Miner uses a novel structure, called utility-list, to store both the utility information about an itemset and the heuristic information for pruning the search space of HUI-Miner. By avoiding the costly generation and utility computation of numerous candidate itemsets, HUI-Miner can efficiently mine high utility itemsets from the utilitylists constructed from a mined database.

Samadi, Pedram,et al[2]In this paper, we consider a smart power infrastructure, where several subscribers share a common energy source. Each subscriber is equipped with an energy consumption controller (ECC) unit as part of its smart meter. Each smart meter is connected to not only the power grid but also a communication infrastructure such as a local area network. This allows two-way communication among smart meters. Considering the importance of energy pricing as an essential tool to develop efficient demand side management strategies, we propose a novel real-time pricing algorithm for the future smart grid. We focus on the interactions between the smart meters and the energy provider through the exchange of control messages which contain subscribers' energy consumption and the real-time price information.

Liu, Ying et al[3]Association rule mining (ARM) identifies frequent itemsets from databases and generates association rules by considering each item in equal value. However, items are actually different in many aspects in a number of real applications, such as retail marketing, network log, etc. The difference between items makes a strong impact on the decision making in these applications. Therefore, traditional ARM cannot meet the demands arising from these applications. By considering the different values of individual items as utilities, utility mining focuses on identifying the itemsets with high utilities. As "downward closure property" doesn't apply to utility mining, the generation of candidate itemsets is the most costly in terms of time and memory space. In this paper, we present a Two-Phase algorithm to efficiently prune down the number of candidates and can precisely obtain the complete set of high utility itemsets. In the first phase, we propose a model that applies the "transaction-weighted downward closure property" on the search space to expedite the identification of candidates. In the second phase, one extra database scan is performed to identify the high utility itemsets. We also parallelize our algorithm on shared memory multi-process architecture using Common Count Partitioned Database (CCPD) strategy.

Tsai, Pauray et al[4]Most of researches on mining high utility itemsets focus on the static transaction database, where all transactions are treated with the same importance and the database can be scanned more than once. With the emergence of new applications, data stream mining has become a significant research topic. In the data stream environment, online data stream mining algorithms are restricted to make only one pass over the data. However, present methods for mining high utility item sets still cannot meet the requirement. In this paper, we propose a single pass algorithm for high utility item set mining based on the weighted sliding window model. The developed algorithm takes advantage of reusing stored information to efficiently discover all the high utility itemsets in data streams.

Wu, Cheng Wei, et al[5]Mining high utility itemsets from databases is an emerging topic in data mining, which refers to the discovery of itemsets with utilities higher than a user-specified minimum utility threshold min_util. Although several studies have been carried out on this topic, setting an appropriate minimum utility threshold is a difficult problem for users. If min_util is set too low, too many high utility itemsets will be generated, which may cause the mining algorithms to become inefficient or even run out of memory. On the other hand, if min_util is set too high, no high utility itemset will be found. Setting appropriate minimum utility thresholds by trial and error is a tedious process for users. In this paper, we address this problem by proposing a new framework named top-k high utility itemset mining, where k is the desired number of high utility itemsets to be mined

III. HIGHTOLLTRANSACTION REDUCTION ALGORITHM

- Group items into higher conceptual groups, e.g. white and brown bread become "bread."
- Reduce the number of scans of the entire database (Apriori needs n+1 scans, where n is the length of the longest pattern)
  - Partition-based apriori
  - Take a subset from the database, generate candidates for frequent itemsets; then confirm the hypothesis on the entire database.

Alternative to Apriori, using fewer scans of database:

- Scan database and find items with frequency greater then or equal to a threshold T
- Order the frequent items in decreasing order
- Construct a tree which has only the root
- Scan database again; for each sample:
  - add the items from the sample to the existing tree, using only the frequent items (i.e. items discovered in step 1.)
  - repeat a. until all samples have been processed
- Enumerate all frequent itemsets by examining the tree: the frequent itemsets are present in those paths for which every node is represented with the frequency ≥ T

IV. PROPOSED FLOWCHART AND RESULTS

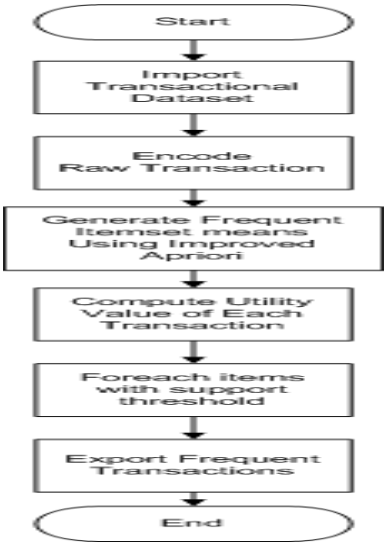

Fig. 1:

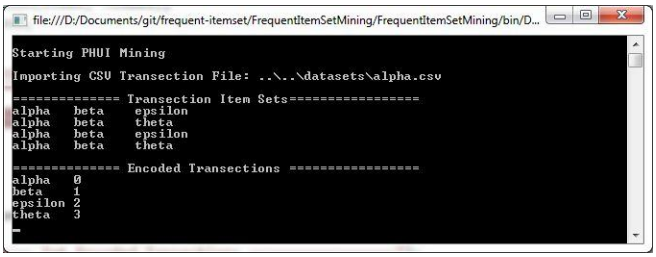

Fig. 2: (encoded transactions)

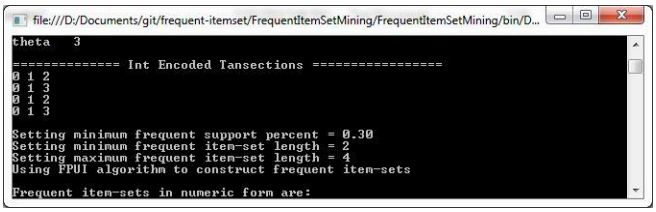

Fig. 3:

Fig. 4:

V. Conclusion

Utility mining is used to find the item-sets in a transactional database with high utility values like profits. Utility mining only focuses on itemsets with high utilities, but the number of rich-enough customers is limited. Traditional association rules mining only concerns the frequency of itemsets, which may not bring large amount of profit. Also Traditional pattern mining algorithms may not find some of the most profitable, high priced patterns, due to their lower support. These algorithms reflect only statistical correlation, but it does not reflect semantic significance of the pattern. This gives reason to develop a mining model to find item-sets, which contributes to business organization with high profit. Hence, utility-based pattern mining technique has evolved and got much popularity in recent time.

Frequent pattern mining has been a focused theme in data mining research for over a decade. Abundant literature has been dedicated to this research and tremendous progress has been made, ranging from efficient and scalable algorithms for frequent itemset mining in transaction databases to numerous research frontiers, such as sequential pattern mining, structured pattern mining, correlation mining, associative classification, and frequent pattern-based clustering, as well as their broad applications. In this research work, we proposed strategies that can, not only, decrease the overestimated utilities of PHUIs but greatly reduce the number of candidates. Different types of both real and synthetic data sets are used in a series of experiments to the performance of the proposed algorithm with state-of-the-art utility mining algorithms. Experimental results show that these algorithms outperform other algorithms substantially in term of execution time, especially when databases contain lots of long transactions or low minimum utility thresholds are set.

The problem of Cost Efficient item-sets mining become one of the most important research area in data mining But all of the existing utility pattern mining algorithms are based on centralized database and today's internet era databases are inherently distributed. We believe that our frequent pattern mining research has substantially broadened the scope of data analysis and will have deep impact on data mining methodologies and applications in the long run. However, there are still some challenging research issues that need to be solved before frequent pattern mining can claim a cornerstone approach in datamining applications.

VI. Future Scope

This research work proposed an algorithm for high utility item set mining, however the algorithm did not reflect the fuzzy degree of quantity and profit level for mined high utility item-sets, which are essential for decision making in various applications like stock control and sales analysis. In future, we will try to apply fuzzy sets theory to the utility mining problem and propose novel methods, for mining fuzzy high utility item-sets. We can also work on the temporal high utility itemsets, which are the itemsets whose support is larger than a pre-specified threshold in current time window of the data stream. Discovery of temporal high utility itemsets is an important process for mining interesting patterns like association rules from data streams

References

[1] Liu, Mengchi, and Junfeng Qu. "Mining high utility itemsets without candidate generation." In Proceedings of the 21st ACM international conference on Information and knowledge management, pp. 55-64.ACM, 2012.

[2] Samadi, Pedram, A-H. Mohsenian-Rad, Robert Schober, Vincent WS Wong, and JuriJatskevich. "Optimal real-time pricing algorithm based on utility maximization for smart grid." In Smart Grid Communications (SmartGridComm), 2010 First IEEE International Conference on, pp. 415-420. IEEE, 2010

[3] Liu, Ying, Wei-keng Liao, and AlokChoudhary. "A fast high utility itemsets mining algorithm."In Proceedings of the 1st international workshop on Utility-based data mining, pp. 90-99.ACM, 2005.

[4] Tsai, Pauray SM. "MINING HIGH UTILITY ITEMSETS IN DATA STREAMS BASED ON THE WEIGHTED SLIDING WINDOWMODEL." International Journal (2014)

[5] Wu, Cheng Wei, Bai-En Shie, Vincent S. Tseng, and Philip S. Yu. "Mining Top-K high utility itemsets." In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 78-86. ACM, 2012

[6] Hong, Tzung-Pei, Cho-Han Lee, and Shyue-Liang Wang. "Mining high average-utility itemsets."In Systems, Man and Cybernetics, 2009.SMC 2009. IEEE International Conference on, pp. 2526-2530. IEEE, 2009.

[7] Cardosa, Michael, Madhukar R. Korupolu, and Aameek Singh. "Shares and utilities based power consolidation in virtualized server environments." In Integrated

Network Management, 2009.IM'09. IFIP/IEEE International Symposium on, pp. 327-334. IEEE, 2009.

[8] Liu, Ying, Wei-keng Liao, and AlokChoudhary. "A two-phase algorithm for fast discovery of high utility itemsets."In Advances in Knowledge Discovery and Data Mining, pp. 689-695.Springer Berlin Heidelberg, 2005.

[9] R. Chithra, and S.Nickolas. "A Tree Based Novel Algorithm for High Utility Itemset Mining", International Journal of Computational Intelligence Research, (2011).

[10] H. Yao, H.J. Hamilton, and C.J. Butz , "A Foundational approach to mining Itemset Utilities From Databases", Third SIAM International Conference on Data Mining, (2004), pp. 482-486.

[11] H. Yao, and H.J. Hamilton, "Mining itemset utilities from Transactional databases, Data & Knowledge Engineering", 59, 2006, pp. 603-626.

[12] R. Agarwal, C. Aggarwal, V.V.V. Prasad, "A tree projection algorithm for generation of Frequent itemsets", Journal of Parallel and Distributed Computing, vol 61, (2001) 350–371.

[13] Han J, Pei J, Yin Y (2000) "Mining frequent patterns without candidate generation" In: Proc of the ACM-SIGMOD int'l conf on management of data, pp 1-12

[14] [13] Pei J, Han J, Mortazavi-Asl B, Pinto H, Chen Q, Dayal U, Hsu MC (2004) "Mining sequential patterns by pattern-growth: the prefix span approach". IEEE Trans Knowl Data Eng 16(10)

[15] Y. Liu, W.K. Liao and A. Choudhary (2005) "A Two Phase algorithm for fast Discovery of high utility itemset", Cheng, D. And Liu, H. (eds) PAKDD, LNCS(LNAI), Vol 3518 pp 689- 695, Springer, Heidelberg.

[16] Adnan M, Alhajj R (2009) DR "FP-tree: disk-resident frequent pattern tree" Appl Intell 30(2):84–97

[17] Agrawal R, Srikant R (1994) "Fast algorithms for mining association rules" In: Proc. of the 20th int'l conf. on very large data bases, pp 487–499

[18] Agrawal R, Srikant R (1995) "Mining sequential patterns" In: Proc of 11th int'l conf on data mining, pp 3–14

[19] Ahmed C. F., Tanbeer S. K., Jeong B.-S., Lee Y.- Koo.: "Efficient Tree Structures for High Utility Pattern Mining in Incremental Databases" Transactions on Knowledge and Data Engineering, Vol. 21, No. 12, pp. 1708 – 1721 (2009)

[20] Ahmed CF, Tanbeer SK, Jeong B-S, Lee Y-K (2011) "HUC-Prune: an efficient candidate pruning technique to mine high utility patterns" Appl Intell 34(2):181–198

[21] Chan R, Yang Q, Shen Y (2003) "Mining high utility itemsets" In: Proc of third IEEE int'l conf on data mining, pp 19–26

[22] Tseng VS, Wu C-W, Shie B-E, Yu PS (2010) "UP-growth: an efficient algorithm for high utility itemsets mining" In: Proc of the 16th ACM SIGKDD conf on knowledge discovery and data mining (KDD'10), pp 253–262

[23] Yao H, Hamilton HJ (2006) "Mining itemset utilities from transaction databases" Data Knowl Eng 59:603–626

[24] Liu Y, Liao W-K, Choudhary A (2005) "A fast high utility itemsets mining algorithm" In: Proc of utility-based data mining Erwin, R.P. Gopalan, and N.R.Achuthan "CTU- Mine: An "Efficient High Utility Itemset Mining Algorithm Using the Pattern growth Approach," Proc. Seventh IEEE Int'l Conf. Computer and Information Technology (CIT '07), 2007,pp. 71-76.

[25] Li, Hua-Fu, Hsin-Yun Huang, Yi-Cheng Chen, Yu-Jiun Liu, and Suh-Yin Lee. "Fast and memory efficient mining of high utility itemsets in data streams." In Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on, pp. 881-886. IEEE, 2008

[26] Davidson, Ian, Kiri L. Wagstaff, and Sugato Basu. Measuring constraint-set utility for partitional clustering algorithms. Springer Berlin Heidelberg, 2006.

[27] Wang, Jing, Ying Liu, Lin Zhou, Yong Shi, and Xingquan Zhu. "Pushing frequency constraint to utility mining model." In Computational Science–ICCS 2007, pp. 685-692. Springer Berlin Heidelberg, 2007.

[28] Yu, Guangzhu, Shihuang Shao, and Xianhui Zeng. "Mining long high utility itemsets in transaction databases." WSEAS Transactions on Information Science & Applications 5, no. 2 (2008): 202-210.

[29] Yu, Guangzhu, Shihuang Shao, and Xianhui Zeng. "Mining long high utility itemsets in transaction databases." WSEAS Transactions on Information Science & Applications 5, no. 2 (2008): 202-210.

