Overview of Itemset Utility Mining and its Applications

Jyothi Pillai
Reader
Bhilai institute of Technology,
Durg-491001, Chhattisgarh, India

O.P.Vyas
Professor
Indian Institute of Information Technology
Allahabad, U.P., India

ABSTRACT

An emerging topic in the field of data mining is Utility Mining. The main objective of Utility Mining is to identify the itemsets with highest utilities, by considering profit, quantity, cost or other user preferences. 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 discovers itemsets that occur frequently. In many real-life applications, high-utility itemsets consist of rare items. Rare itemsets provide useful information in different decision-making domains such as business transactions, medical, security, fraudulent transactions, retail communities. For example, in a supermarket, customers purchase microwave ovens or frying pans rarely as compared to bread, washing powder, soap. But the former transactions yield more profit for the supermarket. Similarly, the high-profit rare itemsets are found to be very useful in many application areas. For example, in medical application, the rare combination of symptoms can provide useful insights for doctors [21]. A retail business may be interested in identifying its most valuable customers i.e. who contribute a major fraction of overall company profit[10]. Several researches about itemset utility mining were proposed. In this paper, a literature survey of various algorithms for high utility rare itemset mining has been presented.

Keywords: Utility Mining, High-utility itemsets, Rare itemsets, Frequent Itemset mining

1. Introduction

1.1 Data Mining

During the last ten years, **Data mining**, also known as knowledge discovery in databases has established its position as a prominent and important research area. The goal of data mining is to extract higher-level hidden information from an abundance of raw data[3]. Data mining has been used in various data domains. Data mining can be regarded as an algorithmic process that takes data as input and yields patterns, such as classification rules, itemsets, association rules, or summaries, as output [11].

Data Mining tasks can be classified into two categories, Descriptive Mining and Predictive Mining. The Descriptive Mining techniques such as Clustering, Association Rule Discovery, Sequential Pattern Discovery, is used to find human-interpretable patterns that describe the data. The Predictive Mining techniques like Classification, Regression, Deviation Detection, use some variables to predict unknown or future values of other variables

1.2 Association rule Mining

Mining Association rules is one of the research problems in data mining [1]. Given a set of transactions where each transaction is a set of items, an association rule is an expression of the form X ⇒ Y, where X and Y are sets of items.

The problem of mining association rules was first introduced in [1] and later broadened in [2], for the case of databases consisting of categorical attributes alone.

Association rule mining (ARM) is a popular technique for finding co-occurrences, correlations, frequent patterns, associations among items in a set of transactions or a database. Rules with confidence and support above user-defined thresholds (minconf and minsup) were found. As data continues to grow and its complexity increases, newer data structures and algorithms are being developed to match this development. Association Rule Mining process can be divided into two steps. The first step involves finding all frequent itemsets (or say large itemsets) in databases. Once the frequent itemsets are found, association rules are generated [6]. ARM is widely used in market-basket analysis. For example, frequent itemsets can be found out by analyzing market basket data and then association rules can be generated by predicting the purchase of other items by conditional probability [1], [2].

1.3 Utility Mining

The traditional ARM approaches consider the utility of the items by its presence in the transaction set. The frequency of itemset is not sufficient to reflect the actual utility of an itemset. For example, the sales manager may not be interested in frequent itemsets that do not generate significant profit.

Recently, one of the most challenging data mining tasks is the mining of high utility itemsets efficiently. Identification of the itemsets with high utilities is called as **Utility Mining.** The utility can be measured in terms of cost, profit or other expressions of user preferences. For example, a computer system may be more profitable than a telephone in terms of profit.

Utility mining model was proposed in [19] to define the utility of itemset. The utility is a measure of how useful or profitable an itemset X is. The utility of an itemset X, i.e., $u(X)$, is the sum of the utilities of itemset X in all the transactions containing X. An itemset X is called a *high utility itemset* if and only if $u(X)$ >= *min_utility*, where *min_utility* is a user-defined minimum utility threshold [11].

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.

1.4 Frequent Itemset Mining

R. Agrawal et al in [1] introduced the concept of frequent itemset mining. **Frequent itemsets** are the itemsets that occur frequently in the transaction data set. The goal of **Frequent Itemset Mining** is to identify all the frequent itemsets in a transaction dataset.

Let $I = \{i1, i2, \ldots, in\}$ be a set of n distinct literals called *items*. An **itemset** is a non-empty set of items. An itemset $X = (i1, i2, \ldots, ik)$ with k items is referred to as k-itemset, A **transaction** T = <TID, $(i1, i2, \ldots, ik)$> consists of a transaction identifier (TID) and a set of items $(i1, i2, \ldots, ik)$, where *ij* Î *I*, $j = 1, 2, \ldots, k$.

The frequency of an itemset X is the probability of X occurring in a transaction T. A frequent itemset is the itemset having frequency support greater a minimum user specified threshold.

1.5 Rare Itemset Mining

The basic Bottleneck of association rule mining is Rare Item Problem. Most approaches to mining association rules implicitly consider the utilities of the itemsets to be equal [6]. The utilities of itemsets may differ. In many applications, some items appear very frequently in the data, while others rarely appear. If frequencies of items vary, two problems encountered – (1) If minsup is set too high, then rules of rare items will not be found (2) To find rules that involve both frequent and rare items, minsup has to be set very low. This may cause *combinatorial explosion.*

In many practical situations, the rare combinations of items in the itemset with high utilities provide very useful insights to the user. **Rare itemsets** are the itemsets that occur infrequently in the transaction data set. In most business applications, frequent itemsets may not generate much profit while rare itemsets may generate a very high profit. Rare itemsets are very important and can be further promoted together because they possess high associations and can bring some acceptable profits [21].

Rare itemsets provide very useful information in the real-life applications such as security, business strategies, biology, medicine and super market shelf-management. For example [16], in the security field, normal behavior is very frequent, whereas abnormal or suspicious behavior is less frequent. Considering a database where the behavior of people in sensitive places such as airports are recorded, if those behaviors are modeled, it is likely to find that normal behaviors can be represented by frequent patterns and suspicious behaviors by rare patterns.

In this paper we have presented a literature survey of the various approaches and algorithms for high-utility mining and rare itemset mining.

2. Literature survey

In the previous section we have introduced the basic concept of Data Mining, Association Rule mining, Utility Mining and Rare Itemset Mining. A brief overview of various algorithms, concepts and techniques defined in different research papers have been given in this section.

The mining of association rules for finding the relationship between data items in large databases is a well studied technique in data mining field with representative methods like ***Apriori*** **[1], [2].** ARM process can be decomposed into two steps. The first step involves finding all frequent itemsets in databases. The second step involves generating association rules from frequent itemsets.

In **[6], Yao et al** defined the problem of utility mining, a 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.

H. Yao et al formalized the semantic significance of utility measures in **[11].** Based on the semantics of applications, the utility-based measures were classified 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.

High utility frequent itemsets contribute the most to a predefined utility, objective function or performance metric[**13**]. For example, From marketing strategy perspective, it is important to identify product combinations that have a significant impact on company's bottom line i.e. having highest revenue generating power[13].

An algorithm for frequent item set mining was presented by **J. Hu et al** in [13**]** that identify high-utility item combinations. In contrast to the traditional association rule and frequent item mining techniques, the goal of the algorithm is to find segments of data, defined through combinations of few items (rules), which satisfy certain conditions as a group and maximize a predefined objective function. The ***highutility 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 [13].

In the paper **[17], H.F. Li** proposed two efficient one-pass algorithms, **MHUI-BIT** and **MHUI-TID**, for mining high utility itemsets from data streams within a transaction-sensitive sliding window. Two effective representations of item information and an extended lexicographical tree-based summary data structure were developed to improve the efficiency of mining high utility itemsets [11].

Liu *et al* proposed **Two-Phase algorithm [8]** for finding high utility itemsets. In the first phase, 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.

A novel method, namely ***THUI*** **(Temporal High Utility Itemsets) –*Mine was*** proposed by **V.S. Tseng et al in [10],** for mining temporal high utility itemsets from data streams efficiently and effectively. The novel contribution of ***THUI-Mine*** is that it can effectively identify the temporal high utility itemsets by generating fewer temporal high transaction-weighted utilization 2-itemsets such that the execution time can be reduced substantially in mining all high utility itemsets in data streams. In this way, the process of discovering all temporal high utility itemsets under all time windows of data streams can be achieved effectively with limited memory space, less candidate itemsets and CPU I/O time. This meets the critical requirements on time and space efficiency for mining data streams. The experimental results show that *THUI-Mine* can discover the temporal high utility itemsets with higher performance and less candidate itemsets compared to other algorithms under various experimental conditions. Moreover, it performs scalable in terms of execution time under large databases. Hence, *THUI-Mine* is promising for mining temporal high utility itemsets in data streams.

G.C.Lan et al proposed a new kind of patterns, named Rare Utility Itemsets in **[21],** which consider not only individual profits and quantities but also common existing periods and branches of items in a multi-database environment. A new mining approach called **TP-RUI-MD (Two-Phase Algorithm for Mining Rare Utility Itemsets in Multiple Databases)** was proposed to efficiently discover rare utility itemsets. The **TP-RUI-MD** algorithm is designed to find the rare-utility itemsets in a multi-database environment. The TP-RUI-MD algorithm uses the level-wise technique for discovering the rare-utility itemsets in a multi-database environment. The TP-RUI-MD algorithm cannot discover all rare-utility itemsets in a multi-database environment, but the TP-RUI-MD algorithm can still discover these rare-utility undiscovered by existing algorithms in a single database environment.

S. Shankar et al, presents a novel algorithm Fast Utility Mining (FUM) in [**19**], which finds all high utility itemsets within the given utility constraint threshold. The authors also suggest a novel method of generating different types of itemsets such as High Utility and High Frequency itemsets (HUHF), High Utility and Low Frequency itemsets (HULF), Low Utility and High Frequency itemsets (LUHF) and Low Utility and Low Frequency itemsets (LULF) using a combination of FUM and Fast Utility Frequent mining (FUFM) algorithms.

In paper **[14], L. Szathmary et al** presented a novel method for computing all rare itemsets by splitting the rare itemset mining task into two steps. The first step is the identification of the minimal rare itemsets. These itemsets jointly act as a minimal generation seed for the entire rare itemset family. In the second step, the minimal rare itemsets are processed in order to restore all rare itemsets. Two algorithms were proposed for the first step: (i) a naive one that relies on an Apriori-style enumeration, Apriori-Rare and (ii) an optimized method that limits the exploration to frequent generators only. The second task is solved by a straightforward procedure. **Apriori-rare** is a modification of the Apriori algorithm used to mine frequent itemsets. Apriori-rare generates a set of all minimal rare generators, also called MRM, that correspond to the itemsets usually pruned by the Apriori algorithm when seeking for frequent itemsets. To retrieve all rare itemsets from minimal rare itemset (mRIs), a prototype algorithm called **"A Rare Itemset Miner Algorithm(Arima)**" was proposed. *Arima* generates the set of all rare itemsets, splits into two sets: the set of rare itemsets having a zero support and the set of rare itemsets with non-zero support. If an itemset is rare then any extension of that itemset will result a rare itemset.

M. Adda et al proposed a framework in **[16],** to represent different categories of interesting patterns and then instantiate it to the specific case of rare patterns. A generic framework was presented to mine patterns based on the Apriori approach. The generalized Apriori framework was instantiated to mine rare itemsets. The resulting approach is Apriori-like and the mine idea behind it is that if the itemset lattice representing the itemset space in classical Apriori approaches is traversed on a bottom-up manner, equivalent properties to the Apriori exploration of frequent itemsets are provided to mine rare itemsets. The Apriori algorithm, called *AfRIM* for *Apriori Rare itemset* to mine rare itemsets was proposed which performs a level-wise descent. The backward traversal method is endowed with a property that leads to prune out potentially non-rare itemsets in the mining process. This includes an anti-monotone property and a level wise exploration of the itemset space.

The authors in **[22]** presents a new foundational approach to temporal weighted itemset utility mining where item utility values are allowed to be dynamic within a specified period of time, unlike traditional approaches where these values are static within those times. Moreover, the approach incorporates a fuzzy model where utilities can assume fuzzy values on the other hand [22]. A Conceptual model has been presented that allows development of an efficient and applicable algorithm to real world data and captures real-life situations in fuzzy temporal weighted utility association rule mining[22].

4. Conclusion

In Data Mining, Association Rule Mining is one of the most important tasks. A large number of efficient algorithms are available for association rule mining, which considers mining of frequent itemsets. But an emerging topic in Data Mining is Utility Mining, which incorporates utility considerations during itemset mining. Utility Mining covers all aspects of economic utility in data mining and helps in detection of rare itemset having high utility. Rare High Utility itemset mining is very beneficial in several real-life applications. In this paper, we have presented a brief overview of various algorithms for high utility rare itemset mining. In the future scope, we will be presenting a comparative study of various algorithms for mining rare high utility itemset.

5. References

[1] R. Agrawal, T. Imielinski and 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 and R. Srikant, 1994, "**Fast Algorithms for Mining Association Rules**", in Proceedings of the 20th International Conference Very Large Databases, pp. 487-499.

[3]Attila Gyenesei, "**Mining Weighted Association Rules for Fuzzy Quantitative Items**", Lecture

notes in Computer Science, Springer, Vol. 1910/2000, pages 187-219, TUCS Technical Report No.346, ISBN 952-12-659-4,ISSN 1239-1891, May 2000.

[4] R. Chan, Q. Yang, Y. D. Shen, "**Mining High utility Itemsets**", In Proc. of the 3rd IEEE Intel. Conf. on Data Mining (ICDM), 2003.

[5] H. Yun, D. Ha, B. Hwang, and K. Ryu. "**Mining association rules on significant rare data using relative support".** Journal of Systems and Software, 67(3):181–191, 2003.

[6] H.Yao, H. J. Hamilton, and C. J. Butz, "**A Foundational Approach to Mining Itemset Utilities from Databases**", Proceedings of the Third SIAM International Conference on Data Mining, Orlando, Florida, pp. 482-486, 2004.

[7] G. Weiss. "**Mining with rarity: a unifying framework",.**SIGKDD Explor. Newsl., 6(1):7–19, 2004.

[8] Liu, Y., Liao, W., and A. Choudhary, A., "**A Fast High Utility Itemsets Mining Algorithm",** In Proceedings of the Utility- Based Data Mining Workshop, August 2005.

[9] Lu, S., Hu, H. and Li, F. 2005. "**Mining weighted association rules. Intelligent Data Analysis"**, 5(3):211–225.

[10]V. S. Tseng, C.J. Chu, T. Liang, "**Efficient Mining of Temporal High Utility Itemsets from Data streams**", Proceedings of Second International Workshop on Utility-Based Data Mining, August 20, 2006

[11]H. Yao, H. Hamilton and L. Geng, "**A Unified Framework for Utilty-Based Measures for Mining Itemsets**", In Proc. of the ACM Intel. Conf. on Utility-Based Data Mining Workshop (UBDM), pp. 28-37, 2006.

[12]A. Erwin, R.P.Gopalan and N. R. Achuthan, 2007, "**A Bottom-up Projection based Algorithm for mining high utility itemsets**", in Proceedings of 2nd Workshop on integrating AI and Data Mining(AIDM 2007)", Australia, Conferences in Research and Practice in Information Technolofy(CRPIT),Vol. 84.

[13] J. Hu, A. Mojsilovic, "**High-utility pattern mining: A method for discovery of high-utility item sets**", Pattern Recognition 40 (2007) 3317 – 3324.

[14] L. Szathmary, A. Napoli, P. Valtchev, "**Towards Rare Itemset Mining**" Proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence, 2007, Volume 1, Pages: 305-312, ISBN ~ ISSN:1082-3409 , 0-7695-3015-X

[15] Kriegel, H-P et al. 2007. "***Future Trends in Data Mining, Data Mining and Knowledge Discovery",*** 15:87–97.

[16] M. Adda, L. Wu, Y. Feng, "**Rare Itemset Mining**", Sixth International conference on Machine Learning and Applications, 2007, pp 73-80.

[17] H.F. Li, H.Y. Huang, Y.Cheng Chen, and Y. Liu and S. Lee, "**Fast and Memory Efficient Mining of High Utility Itemsets in Data Streams**", 2008 Eighth IEEE International Conference on Data Mining.

[18] M. Sulaiman Khan, M. Muyeba, Frans Coenen, 2008. "Fuzzy Weighted Association Rule Mining with Weighted Support and Confidence Framework", to appear in ALSIP (PAKDD),pp. 52-64.

[19]S. Shankar, T.P.Purusothoman, S.Jayanthi and N.Babu, "**A Fast Algorithm for Mining High Utility Itemsets**", Proceedings of IEEE International Advance Computing Conference (IACC 2009), Patiala, India, pages **:** 1459 - 1464

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

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

[22] J. Pillai, O.P. Vyas, S. SoniM. Muyeba "**A Conceptual Approach to Temporal Weighted Itemset Utility Mining**", *2010 International Journal of Computer Applications (0975 - 8887) Volume 1 – No. 28*