International Journal of Computer Science and Engineering Communications,
Volume.5, Issue.2 (2017): Page.1550- 1555
www.ijcsec.com

Survey on Utility Data mining

M. Ganesan
Assistant Professor
Department of Information Technology
Hindusthan College of Engineering and Technology
Coimbatore India
Email: ganeshktg@gmail.com

S.Shankar
Professor
Department of Computer Science and Engineering,
Sri Krishna College of Engineering and Technology Coimbatore India
Email: shanx80@gmail.com

Abstract-- Data Mining is an activity that extracts some new useful information contained in large databases. Traditional data mining methodologies focused largely on detecting the statistical correlations between the items that are more frequent in the transaction databases. Association Rule Mining focuses on existence of an item in a transaction, whether or not it is purchased. The drawbacks of frequent itemset mining leads to consider a utility mining, which allows a user to conveniently express the usefulness of itemsets as utility values and then find itemsets with high utility values. In practice the utility value of an itemset can be profit, popularity, or some other measures of user's preference. There exist several algorithms in literature to mine high utility itemsets. In this paper, a literature survey of various high utility itemset mining algorithms has been presented.

Keywords *--Data Mining; Frequent pattern Mining; Utility Mining*

I INTRODUCTION

The rate of new data generation is enormous. As the advances in electronics and computer technologies have unlimited storage resources, virtually every bit of new data is stored, preserved, and made available. The Internet, an electronic network spread around the hosts an almost unimaginable amount of human-generated data. To get the required knowledge from these data we need some data mining techniques.

A. Data Mining

Data mining denotes the activity of extracting new, valuable and nontrivial information from large volumes of data. Most commonly, the aim is to find patterns or build models using specific algorithms from various scientific disciplines including artificial intelligence, machine learning, and database systems. The data mining tasks can be classified into two categories:

- **Pedictive data mining** where the goal is to build an executable model from data which can be used for classification, prediction or estimation.
- **Descriptive data mining** where the goal is to discover interesting patterns and relationships in data.

DOI: 10.7910/DVN/YXNCHP

B. Frequent Pattern Mining

Frequent itemsets are the items that appear frequently in the transactions. The main goal of frequent itemset mining is to identify all the itemsets in the transaction data set, which are frequently purchased. Item sets are defined as a non empty set of items. If itemset is with k-different items is termed as a k-itemset. For ex{bread, butter, milk} may denoted as a 3-itemset in a supermarket transaction[1].
Let I= {i } be a set of items and D be a transaction database{ } where each transaction T € D is a subset of I. The support or frequency of a pattern X{} is the number of transaction contained the pattern in transactional database.
The Apriori [1], algorithm is the initial solution for the frequent pattern mining problem.To overcomes the problems of Aprori, which generates more candidate sets and require more scans of database FP-Growth has been proposed [2], Uses FP-Tree data structure without any candidate generation and using only two database scans. In the framework of frequent itemsets mining the importance of an item are not considered.

Limitations of Frequent Pattern Mining :
i) The purchase quantities are not taken into account. Thus, an item may only appear once or zero time in a transaction.
ii) All items are viewed as having the same importance, utility of weight. For example, if a customer buys a very expensive bottle of wine or just a piece of bread, it is viewed as being equally important.
Thus, frequent pattern mining may find many frequent patterns that are not interesting. For example, one may find that {bread, milk} is a frequent pattern. However, from a business perspective, this pattern may be uninteresting because it does not generate much profit. Moreover, frequent pattern mining algorithms may miss the rare patterns that generate a high profit such as perhaps {caviar, wine}

C. Utility Mining

Utility mining is 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 as per the user preferences utility can be measured in terms of cost, profit or other expressions. 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. In utility based mining the term utility refers to the quantitative representation of user preference i.e. according to an itemsets utility value is the measurement of the importance of that itemset in the user's perspective.

All Given a finite set of items I = {i1, i2, …, im}. Each item ip ($1 \leq p \leq m$) has a unit profit p(ip). An itemset X is a set of k distinct items {i1, i2, …, ik}, where ij∈I, $1 \leq j \leq k$, and k is the length of X. An itemset with length k is called k-itemset. A transaction database D = {T1, T2, …, Tn} contains a set of transactions, and each transaction Td ($1 \leq d \leq n$) has an unique identifier d, called TID. Each item ip in the transaction Td is associated with a quantity q (ip, Td), that is, the purchased number of ip in Td.[14p9mp9-9-] Consider a simple database with 5 transactions and 7 items.

Example database

Transaction	Item sold in a transaction						
ID	**A**	**B**	**C**	**D**	**E**	**F**	**G**
T1	1	0	1	1	0	0	0
T2	2	0	6	0	2	9	5
T3	1	2	1	6	1	5	0
T4	0	4	3	3	1	0	0
T5	0	2	2	0	1	0	2

Item	A	B	C	D	E	F	G
Profit	5	2	1	2	3	1	1

***Utility of an item*:-** The utility of an item ip in the transaction Td is denoted as u (ip, Td) and defined as p(ip) × q(ip, Td). For example, in Table 1, u ({A}, T1) = 5 × 1 = 5.

***Utility of an itemset*:-** The utility of an item X in Td is denoted as u(X, Td). For example, u ({AC}, T1) =u ({A}, T1) + u({C}, T1) = 5 + 1 = 6. For example, u({AD}) =u({AD}, T1) + u({AD}, T3) = 7 + 17 = 24

***High utility itemset*:-** An itemset is called a high utility itemset if its utility is not less than a user-specified minimum utility threshold which is denoted by Min_util. else; it is called as a low utility itemset.

II LITERATURE SURVEY

Agarwal[1,2] stated 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 and confidence model.

Yao [3] 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 [4] 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.

H.F.Li [7] propose two efficient one pass algorithms MHUI-BIT and MHUI-TID for mining high utility itemsets from data streams within a transaction sensitive sliding window.
Liu et al in [5] proposes a Two-phase algorithm for finding high utility itemsets. Tseng et al in [6] proposes a novel method THUI (Temporal High Utility Itemsets)-Mine for mining temporal high utility itemset 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.

J.Hu [8] presented an algorithm for frequent item set mining that identify high utility item combinations. the goal of the algorithm is to find segments of a data, defined through combinations of some items (rules), which satisfy certain conditions as a group and maximize a predefined objective function In contrast to the traditional association rule and frequent item mining techniques.

S.Shankar [9], presents a novel algorithm Fast Utility Mining (FUM) in, 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

Cheng-Wei Wu [10] presented a novel algorithm with a compact data structure for efficiently discovering high utility itemsets from transactional databases. The UP-Growth is one of the efficient algorithms to generate high utility itemsets depending on construction of a global UP-Tree. In phase I, the framework of UP-Tree follows three steps:
i). Construction of UP-Tree.
ii). Generate PHUIs from UP-Tree.
iii). Identify high utility itemsets using PHUI The construction of global UP-Tree is follows,

- Discarding global unpromising items is to eliminate the low utility items and their utilities from the transaction utilities.

- Discarding global node utilities during global UP-Tree construction. By DGN strategy, node utilities which are nearer to UP-Tree root node are effectively reduced. The PHUI is similar to TWU, which compute all itemsets utility with the help of estimated utility. Finally, identify high utility itemsets.

Erwin [11] 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 itemsets using the pattern growth approach.
Pillai [20] presents a new foundational approach to temporal weighted itemset 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.

Vincent S. [13] proposed a framework for mining Closed High Utility Itemsets (CHUIs). This paper proposed three efficient algorithms named AprioriCH (Apriori-based algorithm for mining High utility closed itemsets), AprioriHC-D (AprioriHC algorithm with Discarding unpromising and isolated items) and CHUD (Closed High Utility Itemset Discovery) to find this representation. To recover all HUIs from the set of CHUIs, authors proposed a method called DAHU (Derive All High Utility Itemsets) and that to without accessing the original database. Authors claimed that this technique achieves a massive reduction in the number of HUIs. AprioriHC-D and AprioriHC both algorithms can't perform well on dense databases when min_utility is low since they suffer from the problem of a large amount of candidates.

III CONCLUSION

In this survey, we analyze from frequent itemset mining to different High Utility Itemset mining algorithms that were proposed by researchers earlier for better development in the field of Data Mining. The multiple algorithms discussed above will be of great use for developing a new improved technique for mining high utility item sets which is efficient and effective. In future we will be developing an algorithm for Mining High Utility Itemsets from Distributed Databases. Most of research on high utility itemset focuses on static databases. With the emergence of the new application, the data processed may be in the continuous dynamic data streams. Because the data in streams come with high speed and are continuous and unbounded, mining result should be generated as fast as possible and make only one pass over a data.

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]. J Han, J Pei, and Y Yin, "Mining Frequent Patterns without Candidate Generation," Proc ACM-SIGMOD Int'l Conf Management of Data, pp. 1-12, 2000.

[3]. H.Yao, H.J.Hamilton, C.J.Butz, A foundation approach to mining itemset utilities from databases,in:Proceedings of the Third SIAM International Conference on Data Mining, Orlando, Florida , 2004,pp.482-486

[4]. H.Yao,H.J.Hamilton ,Mining itemset utilities from transacation databases, in Data and Knowledge Enineering 59(2006) pp.603-626

[5]. Liu. Y, Liao. W,A. Choudhary, A fast high utility itemsets mining algorithm, in: Proceedings of the Utility-Based Data Mining Workshp, August 2005

[6]. V.S.Tseng ,C.J. Chu , T.Liang, Efficient mining of temporal high utility itemsets from data streams, in: Proceedings of Second International Workshop on Utility-Based Data Mining , August 20, 2006

[7]. H.F.Li, H.Y. Huang , Y.Cheng Chen, y. Liu, S.Lee, Fast and memory efficient mining of high utility itemsets in data streams, in :Eigth International Conference of Data Mining 2008

[8]. J.Hu, A. Mojsilovic , High-utility pattern mining :A method for discovery of high-utility ietmsets,in :Pattern Recognition 40(2007) 3317-3324

[9]. S.Shankar, Dr. T .Purusothaman, Kannimuthu.S a novel utility and frequency based itemset mining approach for improving crm in retail business 2010 international journal of computer applications (0975 - 8887) volume 1 – no. 16

[10]. Cheng Wei Wu1, Bai-En Shie1, Philip S. Yu2, Vincent S. Tseng1 Mining Top-K High Utility Itemsets KDD'12, August 12–16, 2012, Beijing, China. Copyright 2012 ACM 978-1-4503-1462-6/12/08

[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.Pillai, O.P.Vyas, S. Soni, M.Muyeba, A conceptual approach to temporal weighted itemset utility mining, in : International Journal of Computer Applications (0975-8887) Volume 1-No.28, 2010

[13]. Vincent S. Tseng, Cheng-Wei Wu, Philippe Fournier-Viger, and Philip S. Yu, "Efficient Algorithms for Mining the Concise and Lossless Representation of High Utility Itemsets", IEEE Transactins on Knowledge and Data Engineering, Vol. 27, No. 3, 2015.

[14]. uangzhou Yu, Shihuang Shao and Xianhui Zeng mining long high utility itemsets in transaction databases wseas transactions on information science & applications issue 2, volume 5, feb. 2008.