IJCSE International Journal of Computer Sciences and Engineering Open Access

Review Paper Volume-4, Issue-12 E-ISSN: 2347-2693

High Utility Pattern Mining – A Deep Review

A. A. Tale[1*], N. R. Wankhade[2]

[1*, 2] Late G. N. Sapkal College of Engineering, Nashik, India

Available online at: www.ijcseonline.org

Received: 17/Nov/2016 Revised: 26/Nov/2016 Accepted: 14/Dec/2016 Published: 31/Dec/2016

Abstract— **The mining high utility pattern is new development in area of data mining. Problem of mining utility pattern with itemset share framework is tricky one as no anti-monotonicity property with interesting measure. Former works on this problem employ a two-phase, candidate generation approach with one exception that is however inefficient and not scalable with large database. This paper reviews former implementation and strategies to mine out high utility pattern in details. We will look ahead some strategies of mining sequential pattern.**

Keywords- **Data Mining, Pattern Mining, High Pattern, Frequent Pattern, Utility Mining.**

I. Introduction

Finding interested patterns in data mining has been an important task, it consists variety of application ranging from genome analysis, condition monitoring, cross-marketing, inventory prediction where interestingness measured [1], [2], [3] play an important role. With frequent pattern mining [4], [5], [6], [7] technique, a pattern is regard interesting if given pattern exceeds user-defined thresholds. For example, mining frequent pattern from online shopping transaction refers to discovery of set of products that are frequently purchased together by customer. However, interest may relate to various factors that can be consider as compared to user-specified terms and occurrences frequency. For example, a company may be interested in discovering combination of product with high profits and revenues which relates to unit profit and purchased quantities of products that are not consider in frequent pattern-mining.

Utility mining [8] recently emerged recently to address the limitation of frequent pattern mining by considering the user expectation or goal as well as the raw data. Utility pattern mining with itemset share framework [9], [10], [11] for example of discovering combinations of products with high profits or revenue, is much harder than other categories of utility pattern mining. Concretely, the interestingness is measured as an anti-monotonicity property and such property can be employed in pruning search space, which is also foundation of all frequent pattern mining.

Most prior to mining high utility pattern with an itemset share framework [10], [12], [13], [14], [15], [16] adopt a two phase, candidate generation approach, that is, first find candidates with high utility pattern in first phase, then scan the raw data one more time to identify high utility patterns from candidates in the second phase. The challenge is most of techniques generate huge number of candidates, which is the scalability and efficiency bottleneck. Such huge number of candidates causes scalability issues not only in first generation but also in first generation. The aim of writing a review paper is to defines in depth review and implementation of high utility mining patterns and further implementation which will help us to implements them.

The paper is organized in section and each section are as follow. Section 2 defines the Utility Mining Problem. Section 3 Survey related works in details. Section 4 discuss various Future Scope and description. Section 5 present the Conclusion.

II. Utility Mining Problem

This section defines the mining utility problem with the itemset share framework that we study.

Let X be the universe of items. Let D be the transaction database $\{t_1, \dots., t_n\}$ where each transaction $t_i \subseteq X$. Each of the transaction item shares a non-zero share. Each transaction of an itemset has independent item weight with respect to transaction given of external utility table. The research problem of finding all high utility patterns are defined as follows. The utility table contains utility of item x in a transaction t. The internal utility is denoted by the share of x in transaction t, whereas the external utility is defined as weight of x independent of any transaction. Hence, the utility of item x is a combination of internal and external utility and utility of function f denotes as non-negative in an application.

Consider the example of a supermarket database transaction. Table I lists the quantities of each production in shopping

[*]Corresponding Author:
A. A. Tale
E-mail: ankittale@hotmail.com, Tel.: +91-77984-95602

transaction where I = $\{a\,, b, c, d, e, f, g\}$ and D = $\{t_{1,}t_2, t_3, t_4, t_5\}$ and Table II list the price of each product.

TABLE I. A SHOPPING TRANSACTION

	Items						
TID	a	b	c	d	e	f	g
t_1	1		1		1		
t_2	6	2	2			5	
t_3	1	1	1	2	6		5
t_4	3	1		4	3		
t_5	2	1		2		2	

TABLE II. UTILITY TABLE

Item	Price
a	1
b	3
c	5
d	2
e	2
f	1
g	1

For transaction $t_2 = \{a, b, c, f\}$ we have $iu(a, t_2) = 6$, $iu(b, t_2) = 2$, $iu(c, t_2) = 2$, $iu(f, t_2) = 5$, $eu(a) = 1, eu(a) = 1, eu(b) = 3$, $eu(c) = 5$, $eu(f) = 1$. Here $u(i, t)$ is product of$iu(i, t)\ and\ eu(i)$. Thus, $u(a, t_2) = 6$, $u(b, t_2) = 6, u(c, t_2) = 10, u(f, t_2) = 5$, and so on.

For Pattern X based on transaction t that is X is subset of t, the utility of X in t denoted as u (X, t) is of transaction that contain utility of each item x in t. A pattern X is high utility pattern, if the utility of X is greater than user-defined minimum thresholds values can be given as

$$\text{HUPset} = \{\, X \mid X \subseteq I, u(X) > minU \,\} \qquad (1)$$

Here HUPSet is short form for high utility pattern and minU denotes the minimum utility thresholds value. In the above example, the market-manager wants to know every combination of products with sales revenues no less than 30 that is minU=30. Since TS ({a, b}) = $\{t_2, t_3, t_4, t_5\}$, we have u ({a, b}) = $u(\{a, b\}, t_2) + u(\{a, b\}, t_3) + u(\{a, b\}, t_4) + u(\{a, b\}, t_5) = u(a, t_2) + u(b, t_2) + u(a, t_3) + u(b, t_3) + u(a, t_4) + u(b, t_4) + u(a, t_5) + u(b, t_5)$ = 27. Similarly, u ({a, c}) = 28, u {(b, c)} = 24, u ({a, b, c}) = 31, u ({a, b, c, d}) =13 and so on. Therefore, HUPSet = {{a, b, c}, {a, b, d}, {a, d, e}, {a. b, d, e}, {b, d, e}, {d, e}, {a, b, c, d, e, g}}.

This is how the high utility patterns are found and only those pattern are which satisfied as user specified thresholds values are selected.

III. Related Work

High Utility Pattern Mining problem is closely related to frequent pattern mining, including constraint-based mining. In this section we briefly review prior works on both mining techniques.

A. Frequent Pattern Mining

Frequent Pattern Mining was first proposed by Agarwal [4], which discover all pattern that are no less than user specified threshold value .It employ anti-monotonicity property, the support of superset of pattern is no more than the support of the pattern. He also design mining algorithm and framework based on level-wise exploration for increasing order of itemsets size. The frequent pattern mining itemsets mining because the tree can be explored in variety of strategies such as depth-first search, breadth-first search and hybrid search. One advantage of using breadth first strategy is that level-wise pruning can be used, which is not possible with other strategies. The challenged of frequent pattern mining is the large number of redundant patterns are often mined out.

Apriori by Agarwal and Srikant [5] is very famous algorithm mining frequent pattern. The Apriori algorithm approach uses a join-based algorithm where all frequents itemsets of length. The main objective which is used for Apriori algorithm is that every subsets of frequent pattern is also frequent, which scan the disk-resident database as many times as the maximum length of frequent pattern. FP-growth by Han [17] approaches combine suffix-based pattern exploration with compressed representation of projected database for more efficient counting. It's based on FP-Tree viewed as trie data-structure of the transaction database of frequent items. Éclat by Zaki [18] is famous hybrid algorithm. Keeps database or partition in memory and work in breadth and depth first search techniques.

B. Constraint-Based Mining

Constraint based mining is a milestone in evolving from frequent pattern mining to utility mining. The constraint based mining to enumerate all patterns that satisfy some constraint. Work in this area is mostly focus on how to push constraint for frequent pattern mining. Pei [19] study constraint which cannot be handled with existing theory and techniques in frequent pattern mining. The develop notion of convertible constraint and systematically analyse, classify, and characterize the class and also implement techniques which enables them to the recently developed FP-Growth algorithm for frequent itemset mining.

Bucila [20] considered mining pattern that satisfy a both monotone and anti-monotone constraints, they proposed DualMiner algorithm that efficiently prune its search space using both anti-monotone and monotone constraints. Bonchi [21] invented ExAnte, a simple yet effective pre-processing techniques for frequent pattern mining. ExAnte property exploits constraints to dramatically reduce the analysed data to containing pattern of interest. The data reduction in turn introduce strong reduction of the candidate pattern search space.

De Raedt [22] investigate how standard constraint programming techniques can be applied to constraint-based mining problem with constraint that are monotone, anti-monotone, and convertible. Bayardo and Agarwal [23] and Morishita and Sese [24] proposed techniques of pruning upper bounds when the constraints are monotone, anti-monotone nor convertible.

C. Categories of Utility Mining

Interestingness measures can be classified as objective, subjective and semantics measure [1]. Objective measures refers to confidence are based only on the data; Subjective measures [25], [26] such as unexpectedness or novelty, or taken into account the user knowledge domain, Semantic measure [3] also known as utilities, consider the data as well as user's expectation.

Hilderman [27] proposed the share-confidence framework for knowledge discovery from database which address the problem of mining itemsets from market basket data. An algorithm for classifying itemsets based upon characteristic attributes extracted from census or lifestyle data. Yao [10], [28] proposed approach permits users to quantify their preference concerning the usefulness of itemsets using utility values. Cai [29] proposed the mining on weighted itemset mining. Lin et.al [30] borrow the term to denote a minor semantic addition to the well-known association rule. They consider the addition of numerical values at the attributes values.

Lu [31] proposed vertical and weighted association rule and also assigned weighted to each item to identify important items. They also present an algorithm to handle problem of mining mixed weighted association rule. Shen [32] and Chan [33] proposed object-oriented utility-based association mining that explicitly models the association of specific "Pattern → Objective" where Pattern is logic-expression asserting objective-attributes and Objective assigning some positive utility.

D. Algorithms with the Itemset Share Framework

As utility mining with itemset share framework is neither monotone, anti-monotone nor convertible most prior algorithm resort to an interim measure proposed by Liu [15] proposed anti-monotonicity property with TWU present Two-phase algorithm to efficiently prune down the candidate and can precisely hold the high utility itemsets and one more database scan is perform to identity the high utility itemsets.

Li [14] proposed the Isolated Items Discarding Strategy which can be applied existing level-wise utility mining method to reduce candidate and to improve performance An isolated item are terms that does not contain any length-k candidate, and hence will occur in any candidate. Any multi-pass, level-wise tree structure can employs IIDS to reduce number of candidate to significant number.

Lan [34] proposed an efficient utility mining approach that adopts an indexing mechanism to speed up the execution and reduce the memory requirement in mining process. The indexing mechanism can imitate the traditional projection sub-database mining. Erwin [13] proposed an algorithm that uses TWU with pattern growth based on compact utility pattern tree data structure. In this the algorithm first identifies the large TWU items first identifies large TWU items in transaction database and if the dataset really small then it forms CUP- Tree for mining high utility patterns

Ahmed [12] proposed three novel tree structure to perform incremental and interactive HUP mining. In first three structure Incremental HUP Lexicographic Tree is arranged according to item's lexicographic order. It can capture incremental data without any restructuring operation. The second tree structure is IHUP Transaction-Frequency which obtains compact forms by arranging the items according to their frequencies and third algorithm is IHUP- Transaction-Weighted Utilization Tree is designed based on the TWU values of item in descending orders.

Tseng [16] proposed algorithms namely utility pattern growth (UP-Growth) and UP-Growth+ for mining HUP with a set of effective strategies for pruning candidates itemsets. It maintain a tree based data structure such that candidate itemsets can be generated efficiently with two scan of database. Yun [35] and Dawar and Goyal [36] improved UPGrowth, by pruning more candidates, while inherent issues of two-phase approach remains.

Liu and Qu [37] simultaneously and independently, proposed to mine high utility patterns without candidate's generation the HUIMiner algorithm by employs a vertical data structure to represent utility information, which employs inefficient join operation and is not scalable. HUIMiner is even less efficient than improved UPGrowth algorithm mining large datasets

Liu and Wang [38] proposed a novel way of mining high utility pattern in single phase without generating candidates in this system they design a look-ahead strategy, a linear data structure CAUL [38] which keep original high utility patterns.

IV. Discussion

All the tree structure implements for single system and main memory based system. In general applying particular techniques in a single system are not scalable and capable of handling the bottleneck in system and mostly they are not good performs at sequential pattern mining, parallel and distributed algorithm. To overcome working on parallel and sequential pattern mining techniques we will discuss some techniques that will helpful in mining utility pattern among them.

The first way, mining high utility sequential pattern by mining pattern in incremental using IncHUSP-Miner[39] algorithm, here we can introduce tight upper bound of the utility sequence called as TSU, and then design novel data structure called as candidate pattern tree, to maintain sequence whose TSU values are greater than or equal mining utility thresholds. The second way, with distributed and parallel high utility sequential pattern by defining BigHUSP [40] an algorithm which help in HUSPs efficiently. In this they also proposed a pruning strategies to minimize search space in distributed environment which helps in decreasing computational and communication cost while maintaining correctness.

Another way by developing a dynamic programming [41] approach to mine probabilistic frequent sequential pattern distributed computing based on Spark. We can a memory efficient distributed DP approach and use extends prefix-tree to intermediate results efficiently. This are various approach way in which we can design for sequential pattern mining and implementing with parallel and distributed algorithm.

Conclusion

In this paper, we discussed different strategy implemented mostly for disk resident and single main memory implementation which are used for mining high utility itemsets. In future we can improves further by re-modelling and re-implemented for designing for parallel and distributed pattern mining process.

Acknowledgment

It give me pleasure in presenting the survey paper for project Stage-I on “Mining High Utility Patterns”. On this occasion I would like to take this opportunity to thank my internal guide Prof. N. R. Wankhade for giving me all help and guidance I needed for writing this paper. I am really grateful to them for their kind support their valuable support were very helpful.

References

[1] L. Geng and H. J. Hamilton, “Interestingness measure for data mining: A survey,” ACM Comput. Survey, Volume-**38**, No. **3**, Page No **9**, **2006**

[2] A. Silberschatz and A. Tuzhilin, “On subjective measures of interestingness in knowledge discovery,” in Proc. ACM 1st Int. Conf. Knowl. Discovery Data Mining, Page No **(275-281)**, **1995**

[3] H. Yao, H. J. Hamilton, and L. Geng, “A unified framework for utility-based measures for mining itemsets,” in Proc. ACM SIGKDD 2nd Workshop Utility-Based Data Mining, pp **(28–37)**, **2006**.

[4] R. Agrawal, T. Imielinski, and A. Swami, “Mining association rules between sets of items in large databases,” in Proc. ACM SIGMOD Int. Conf. Manage. Data, pp. **(207–216)**, **1993**.

[5] R. Agrawal and R. Srikant, “Fast algorithms for mining association rules,” in Proc. 20th Int. Conf. Very Large Databases, pp. **(487–499)**, **1994**.

[6] J. Han, J. Pei, and Y. Yin, “Mining frequent patterns without candidate generation,” in Proc. ACM SIGMOD Int. Conf. Manage. Data, pp. **(1–12)**, **2000**.

[7] M. J. Zaki, “Scalable algorithms for association mining,” IEEE Trans. Knowl. Data Eng., Volume. **12**, no. **3**, pp. **(372–390)**, May/Jun. **2000**.

[8] H. Yao, H. J. Hamilton, and L. Geng, “A unified framework for utility-based measures for mining itemsets,” in Proc. ACM SIGKDD 2nd Workshop Utility-Based Data Mining, pp. **(28–37)**, **2006**.

[9] R. J. Hilderman, C. L. Carter, H. J. Hamilton, and N. Cercone, “Mining market basket data using share measures and characterized itemsets,” in Proc. PAKDD, pp. **(72–86)**, **1998**.

[10] H. Yao and H. J. Hamilton, “Mining itemset utilities from transaction databases,” Data Knowl. Eng., Volume- **59**, No- **3**, Page No **(603–626)**, **2006**.

[11] H. Yao, H. J. Hamilton, and C. J. Butz, “A foundational approach to mining itemset utilities from databases,” in Proc. SIAM Int. Conf. Data Mining, pp. **(482–486)**, **2004**.

[12] C. F. Ahmed, S. K. Tanbeer, B.-S. Jeong, and Y.-K. Lee, “Efficient tree structures for high utility pattern mining in incremental databases,” IEEE Trans. Knowl. Data Eng., Volume-**21**, No-**12**, Page No – **(1708– 1721)**, Dec. **2009**.

[13] A. Erwin, R. P. Gopalan, and N. R. Achuthan, “Efficient mining of high utility itemsets from large datasets,” in Proc. 12th Pacific-Asia Conf. Adv. Knowl. Discovery Data Mining, Page No- **(554–561)**, **2008**.

[14] Y.-C. Li, J.-S. Yeh, and C.-C. Chang, “Isolated items discarding strategy for discovering high utility itemsets,” Data Knowl. Eng., Volume-**64**, No. **1**, Page No. **(198–217)**, **2008**.

[15] Y. Liu, W. Liao, and A. Choudhary, “A fast high utility itemsets mining algorithm,” in Proc. Utility-Based Data Mining Workshop SIGKDD, Page No. **(253–262)**, **2005**.

[16] V. S. Tseng, B.-E. Shie, C.-W. Wu, and P. S. Yu, “Efficient algorithms for mining high utility itemsets from transactional databases,” IEEE Trans. Knowl. Data Eng., Volume- **25**, No-**8**, Page No – **(1772–1786)**, Aug. **2013**.

[17] J. Han, J. Pei, and Y. Yin, “Mining frequent patterns without candidate generation,” in Proc. ACM SIGMOD Int. Conf. Manage. Data, Page No- **(1–12)**, **2000**.

[18] M. J. Zaki, “Scalable algorithms for association mining,” IEEE Trans. Knowl. Data Eng., Volume - **12**, No- **3**, Page No – **(372–390)**, May/Jun. **2000**.

[19] J. Pei, J. Han, and V. Lakshmanan, “Pushing convertible constraints in frequent itemset mining,” Data Mining Knowl. Discovery, Volume - **8**, No - **3**, Page No – **(227–252)**, **2004**.

[20] C. Bucila, J. Gehrke, D. Kifer, and W. M. White, “Dualminer: A dual-pruning algorithm for itemsets with constraints,” Data Mining Knowl. Discovery, Volume - **7**, No - **3**, Page No – (**241–272)**, **2003**.

[21] F. Bonchi, F. Giannotti, A. Mazzanti, and D. Pedreschi, "ExAnte: A preprocessing method for frequent-pattern mining," IEEE Intell. Syst., Volume - **20**, No - **3**, Page No – **(25–31)**, May/Jun. **2005.**

[22] L. De Raedt, T. Guns, and S. Nijssen, "Constraint programming for itemset mining," in Proc. ACM SIGKDD, Page No – **(204–212)**, **2008.**

[23] R. Bayardo and R. Agrawal, "Mining the most interesting rules," in Proc. 5th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, Page No- **(145–154)**, **1999**.

[24] S. Morishita and J. Sese, "Traversing itemset lattice with statistical metric pruning," in Proc. 19th ACM Symp. Principles Database Syst., Page No - **(226–236)**, **2000**.

[25] R. J. Hilderman and H. J. Hamilton, "Measuring the interestingness of discovered knowledge: A principled approach," Intell. Data Anal., Volume -**7**, No- **4**, Page No- **(347–382)**, **2003**.

[26] P. N. Tan, V. Kumar, and J. Srivastava,, "Selecting the right objective measure for association analysis," Inf. Syst., Volume - **29**, No - **4**, Page No – **(293–313)**, **2004**.

[27] R. J. Hilderman, C. L. Carter, H. J. Hamilton, and N. Cercone, "Mining market basket data using share measures and characterized itemsets," in Proc. PAKDD, Page No -**(72–86)**, **1998**.

[28] H. Yao, H. J. Hamilton, and C. J. Butz, "A foundational approach to mining itemset utilities from databases," in Proc. SIAM Int. Conf. Data Mining, Page No – **(482–486)**, **2004**.

[29] C. H. Cai, A. W. C. Fu, C. H. Cheng, and W. W. Kwong, "Mining association rules with weighted items," in Proc. Int. Database Eng. Appl. Symp., Page No – **(68–77)**, **1998**.

[30] T. Y. Lin, Y. Y. Yao, and E. Louie, "Value added association rules," in Proc. 6th Pacific-Asia Conf. Adv. Knowl. Discovery Data Mining, Page No - **(328–333)**, **2002**.

[31] S. Lu, H. Hu, and F. Li, "Mining weighted association rules," Intell. Data Anal., Volume- **5**, No - **3**, Page No – **(211–225)**, **2001**.

[32] Y. Shen, Q. Yang, and Z. Zhang, "Objective-oriented utility-based association mining," in Proc. IEEE Int. Conf. Data Mining, Page No – **(426–433)**, **2002**.

[33] R. Chan, Q. Yang, and Y. Shen, "Mining high utility itemsets," in Proc. Int. Conf. Data Mining, Page No –**(19–26)**, **2003**.

[34] G.-C. Lan, T.-P. Hong, and V. S. Tseng, "An efficient projectionbased indexing approach for mining high utility itemsets," Knowl. Inf. Syst., Volume - **38**, No - **1**, Page No – **(85–107)**, **2014**.

[35] U. Yun, H. Ryang, and K. H. Ryu, "High utility itemset mining with techniques for reducing overestimated utilities and pruning candidates," Expert Syst. Appl., Volume - **41**, No - **8**, Page No – **(3861–3878)**, **2014**.

[36] S. Dawar and V. Goyal, "UP-Hist tree: An efficient data structure for mining high utility patterns from transaction databases," in Proc. 19th Int. Database Eng. Appl. Symp, Page No – **(56–61), 2015.**

[37] M. Liu and J. Qu, "Mining high utility itemsets without candidate generation," in Proc. ACM Conf. Inf. Knowl. Manage. ,Page No – **(55–64)** , **2012**.

[38] J. Liu, Ke Wang, and C. M. Fung, "Mining High Utility Pattern in One Phase without Generating Candidate," IEEE Transaction on Knowledge and Data Engineering, Volume - **28**, No – **5,** Page No – **(1245–1257)**, May- **2016.**

[39] Jun-Zhe. Wang and Juin-Long Huang, "Itemset mining of High Utility Sequential Pattern in Incremental Database," ACM. CIKM'16, Indianapolis, IN, USA, pp – **(1245–1257)**, October 24th- 28th, **2016.**

[40] M. Zihayat, Z. Zhenhua Hu, A. An and Yonggang Hu "Distributedand Parallel High Utility Sequential Pattern Mining," Technical Report EECS – **2016-03**, April 12th, **2016**

[41] Jiaqi Ge and Yuni Xia, "Distributed Sequential Pattern Mining in Large Scale Uncertain Databases," PAKDD Springer International Publishing Switzerland, pp – **(17 -29) , 2016**

Author's Profile

Mr. Ankit A Tale pursed Bachelor of Computer Engineering from University of Pune, India in 2014. He is currently pursuing Master of Computer Engineering from University of Pune, India and currently working as Developer in VirtueTech Solution,since 2015. He has published one research papers in reputed international journals and it's also available online. His main research work focuses on Algorithms, Machine Learning, Artifical Intellegence, Big Data Analytics, Data Mining, IoT based education. He has 2 years of developer experience.

Mr N R Wankhade pursed Bachelor of Computer Engineering from University of Amravati in year 2009. He is currently pursuing Ph.D. and currently working as Professor in Department of Computer Engineering, University of Pune, India since 2005. He is a member of IEEE & IEEE computer society since 2013. He has published more than 20 research papers in reputed international journals including Thomson Reuters (SCI & Web of Science) and conferences including IEEE and it's also available online. His main research work focuses on Cryptography Algorithms, Network Security, Cloud Security and Privacy, IoT and Computational Intelligence based education. He has 20 years of teaching experience and 4 years of Research Experience.