

Systematic Survey on High-Utility Itemset Mining

Mr. Sandipkumar C. Sagare¹ and Prof. Dr. Dattatraya V. Kodavade²

¹⁻²D.K.T.E. Society's Textile and Engineering Institute, Ichalkaranji, Kolhapur, India

Email: sagaresandip999@gmail.com, dvkodavade@gmail.com

Abstract—High utility itemset mining is a popular research field, which consists of discovering the itemsets having a high profit or importance in a database. The utility of an itemset can be calculated based on various criteria like profit, frequency of itemset, and weights etc. High utility itemset is a set of itemsets having a higher importance for users according to utility function. The effort of mining high utility itemset actually generalize the problem of frequent itemset mining as it also considers item quantities and weights. This paper reviews the variety of algorithms developed for discovering high-utility itemsets. It also presents an analytical comparison based on datasets used; data structures used; approach and lastly, it discusses future research opportunities.

Index Terms— High-utility, Mining, Pattern, Itemset.

I. INTRODUCTION

Data mining is the basic and very much important task for getting useful information from available data. Many businesses use this extracted information for their product and business analysis. One of the famous data mining tasks is the frequent Itemset mining (FIM) [1], [2], [3], [4], [5], [6], [7] and [8]. Processes for finding frequent itemsets usually assume a threshold for comparing it with the support of Itemset. The itemsets qualifying after the comparison are termed frequent itemsets. During this process of discovering frequent itemsets, all items in the transaction database are considered to be equally important and also, they can appear at most once per transaction. Three notable limitations of frequent Itemset mining are as follows. First, quantities of items purchased in the transactions are neglected. So, purchasing some units of an item is considered equal as purchasing a single unit. Second, All items in transactions are given equal importance. But in reality, several items can be more important to users. Third, frequent itemsets may not be more interesting to users. E.g. In the market basket analysis, profit gained may be more important than the frequency of selling. These limitations of frequent Itemset mining algorithms are addressed by High-Utility Itemset Mining (HUIM) [9] to [13].

TABLE I. A TRANSACTION DATABASE

Tid	Transactions
T1	(A, 1), (C, 1), (D, 3)
T2	(A, 2), (C, 6), (E, 2), (F, 4)
T3	(A, 1), (B, 2), (C, 3), (D, 3), (E, 1)
T4	(B, 2), (C, 2), (E, 1), (F, 2)
T5	(A, 2), (F, 5)

TABLE II. UTILITIES OF ITEMS

Item Name	A	B	C	D	E	F
Utility	3	6	5	8	5	5

II. LITERATURE REVIEW

R. Agrawal, R. Srikant [1] considered the problem of discovering association rules between items in a large database of sales transactions. They presented two new algorithms for solving the problem. The first step of these algorithms was frequent itemset mining. Limitation of this work was that it didn't consider quantities of items purchased in the transactions.

J. Han, J. Pei, Y. Yin, R. Mao [2] proposed a novel structure called frequent-pattern tree (FP-tree) structure, storing information about the frequent patterns, and also developed an efficient method for FP-tree-based mining. It was scalable and efficient short/long frequent patterns and was proposed to be faster than the Apriori algorithm. However it had the same limitation of not considering the quantities of purchased items.

T. Uno, M. Kiyomi, H. Arimura [4] discovered a fast implementation of LCM for enumerating frequent closed itemsets, that was based upon prefix-preserving closure extension. They implemented LCMfreq and LCMmax for discovering frequent itemsets as well as maximal frequent itemsets by making modifications in LCM. It worked better for sparse datasets.

Z. Farzanyar, M. Kangavari, N. Cercone [5] proposed Max-FISM (Maximal-Frequent Itemsets Mining) algorithm to mine recent maximal frequent itemsets out of the high-speed transactions stream within window called a sliding window. In this algorithm, whenever a transaction got inserted in the given window only the maximum itemset from that window was inserted into summary data structure based on prefix tree called Max-Set. Finally, from current Max-Set, the set of recent maximal frequent itemsets was discovered.

Fournier-Viger, P., Lin, J.C.-W., Dinh, T., Le, H.B. [8] introduced the concept of correlation in the mining of high-utility itemsets to find highly correlated profit giving itemsets FCHM algorithm (Fast Correlated high-utility itemset Miner) was used for this purpose.

D. Kim, U. Yun [9] discussed the high utility itemset mining along with its characteristics to find more meaningful itemsets as compared to other approaches by considering the utility of each item in a given database. They worked on mining high utility itemsets over the data streams which gets continually generated at a rapid rates. They proposed a tree-based algorithm that mined recent high utility itemsets over data streams. They assigned larger weights for recent data as compared to those of older ones. The algorithm the utility information updated regularly in its tree data structure.

P. Fournier-Viger, J.C.W. Lin, T. Dinh, H.B. Le [10] proposed to link the concept of correlation in high-utility itemset mining to discover profitable itemsets that are highly correlated. For this they used the bond measure. An efficient algorithm named FCHM (Fast Correlated high-utility itemset Miner) was developed to efficiently find correlated high-utility itemsets. They proved with experimental results that FCHM was highly-efficient and could prune a huge amount of weakly correlated HUIs.

U. Yun, D. Kim, E. Yoon, H. Fujita [11] introduced Utility pattern mining (UPM) for evaluation of the usefulness of patterns. They employed the high average utility pattern mining (HAUPM) technique, which was one of the UPM approaches as well as discovered interesting patterns out of which these items have more meaningful relations between each other using utility measure. Their approach also adopted damped window based model for finding useful patterns in stream based environments. They performed various experiments on real-life datasets to show effectiveness of their approach.

Philippe Fournier-Viger, Yimin Zhang, Jerry Chun-Wei Lin, Hamido Fujita, Yun Sing Koh [12] identified the problem of mining Local High-Utility Itemsets (LHUIs), and then extended that to discover Peak High-Utility Itemsets (PHUIs), and third problem was proposed for reducing redundancy in the set of discovered PHUIs, that is to find non-redundant peak high utility itemsets. Three algorithms named LHUI-Miner, PHUI-Miner and NPHUI-Miner were designed for efficiently find LHUIs, PHUIs as well as NPHUIs.

III. ANALYSIS OF DIFFERENT TECHNIQUES FOR HIGH-UTILITY ITEMSET MINING

Sr.No	Algorithm	Phases	Type of Search	Dataset type & Data Structure used	Approach	Evaluation of Performance
1	HUI-Miner	1	DFS	Vertical (utility-list)	Skips generate & test approach of candidate itemsets.	Inefficient joins, not scalable.
2	FHM	1	DFS	Vertical (utility-list)	HUI-Miner Extension,	less joins, so faster than HUI-Miner

3	mHUI-Miner	1	DFS	Vertical (utility-list)	Global tree.	Superior for sparse datasets and competitive for dense datasets.
4	Two-Phase	2	BFS	Horizontal	Apriori based, overestimation of utilities in phase 1, then filters in phase 2.	Inefficient, generates Vast number of candidates in phase 1.
5	UP-Growth	2	DFS	Horizontal (prefix-tree)	FP-Growth based, constructs UP-tree	Fewer candidates. Efficient for dense datasets.
6	D2HUP	1	DFS	Vertical (hyper linked list-chain of accurate utility list)	Maintains CAUL-chain for utility list	faster than HUI-Miner
7	HUP-Miner	1	DFS	Vertical (partitioned utility-list)	HUI-Miner Extension, pruning& partitioning used.	faster than HUI-Miner
8	EFIM	1	DFS	Horizontal (merging)	Linear time search is possible	Low memory due to projected database and merging.

IV. CONCLUSION

High utility itemset mining is an progressive area of research. The utility information is interesting for various decision-making purposes in areas like retail stores, banks, stock market etc. This paper has presented a systematic survey of various useful algorithms in the domain of high-utility itemset mining, which can be utilized for developing more useful and efficient techniques in this area. Various extensions are also possible by applying constraints or using parallel approaches that opens many research opportunities. The future work can incorporate designing the frameworks to for high-utility itemset mining on real-life data sets.

REFERENCES

- [1] R. Agrawal, R. Srikant, "Fast algorithms for mining association rules in large databases", in: Proc. 20th Int. Conf. Very Large Databases, 1994, pp. 4 87–4 99. Morgan Kaufmann, Santiago de Chile.
- [2] J. Han, J. Pei, Y. Yin, R. Mao, "Mining frequent patterns without candidate generation: a frequent-pattern tree approach", Data Min. Knowl. Discov. 8 (1) (2004) 53–87.
- [3] J. Han, J. Pei, Y. Yin, R. Mao, "Mining frequent patterns without candidate generation: a frequent-pattern tree approach", Data Min. Knowl. Discov. 8 (1) (2004) 53–87.
- [4] T. Uno, M. Kiyomi, H. Arimura, "LCM ver. 2: efficient mining algorithms for frequent/closed/maximal itemsets", in: Proc. ICDM'04 Workshop on Frequent Itemset Mining Implementations, Vol. 126, CEUR, Brighton, 2004.
- [5] Z. Farzanyar, M. Kangavari, N. Cercone, "Max-FISM: mining (recently) maximal frequent itemsets over data streams using the sliding window model", Comput. Math. Appl. 64 (6) (2012) 1706–1718.
- [6] P. Fournier-Viger, J.C.W. Lin, B. Vo, T.T. Chi, J. Zhang, H.B. Le, "A survey of itemset mining", WIREs Data Min. Knowl. Discovery e1207 (2017), doi:10.1002/ wdm.1207.
- [7] P. Fournier-Viger, J.C.W. Lin, R.U. Kiran, Y.S. Koh, R. Thomas, "A survey of sequential pattern mining", Data Sci. Pattern Recognit. (DSPR) 1 (1) (2017) 54–77.
- [8] Fournier-Viger, P., Lin, J.C.-W., Dinh, T., Le, H.B.: Mining correlated high-utility itemsets using the bond measure. In: Proceedings of the International Conference Hybrid Artificial Intelligence Systems, pp. 53–65. Springer (2016).
- [9] D. Kim, U. Yun, "Mining high utility itemsets based on the time decaying model", Intell. Data Anal. 20 (5) (2016) 1157–1180.
- [10] P. Fournier-Viger, J.C.W. Lin, T. Dinh, H.B. Le, "Mining correlated high-utility itemsets using the bond measure", in: Proc. Intern. Conf. Hybrid Artificial Intelligence Systems, Springer Seville, 2016, pp. 53–65.
- [11] U. Yun, D. Kim, E. Yoon, H. Fujita, "Damped window based high average utility pattern mining over data streams", Knowl. Based Syst. 144 (15) (2018) 188–205.
- [12] Philippe Fournier-Viger, Yimin Zhang , Jerry Chun-Wei Lin , Hamido Fujita , Yun Sing Koh, "Mining local and peak high utility itemsets", et al. / Information Sciences 481 (2019) 344–367.



A) Mr. Sandipkumar C. Sagare is pursuing Ph.D. in Computer Science and Engineering from Shivaji University, Kolhapur. He is currently working as Assistant Professor in the Computer Science and Engineering Department of D.K.T.E. Society's Textile and Engineering Institute Ichalkaranji, India. He received the B.E., M.E. in Computer Science and Engineering from the Shivaji University, Kolhapur, Maharashtra, India. His research interests include high-utility itemset mining, sequential rule mining, data analytics, and machine learning. He is also nominee member of Computer Society of India from D.K.T.E. Society's Textile and Engineering Institute Ichalkaranji.



B) Dr. Dattatraya V. Kodavade is presently working as Professor and Head at Computer Science & Engineering department at D.K.T.E. Society's Textile and Engineering Institute Ichalkaranji, India. He has received B.E., ME. & Ph.D. degree from Shivaji University, Kolhapur. He is having 27 years of teaching experience and has published more than 20 papers in International Journals. He is working as reviewer for International Journal on Super Computing, Springer USA. He is also Editor in Chief for International Journal on Knowledge Based Computer Systems, New Delhi. He is working as member of technical program committee for many National and International reputed conference in computing. His areas of research are artificial Intelligence, Machine learning, IoT, Robotics and High Performance computing