DOI: https://doi.org/10.46632/daai/2/1/10

Data Analytics and Artificial Intelligence
Vol: 2(1), 2022
REST Publisher; ISBN: 978-81-948459-4-2
Website: http://restpublisher.com/book-series/data-analytics-and-artificial-intelligence/

An Enhanced Fast – High Utility Item set Mining Method for Large Datasets

K. Subba Reddy, N. Harshada, B. Suhasini, M. Bhanu Priya, G. Roopaswini
Rajeev Gandhi Memorial College of Engineering & Technology, Nandyal, Kurnool, Andhra Pradesh, India
*Corresponding author Email: mrsubbareddy@yahoo.com

Abstract. High Utility Itemset mining is considered one of the critical and challenging problems in data mining. The existing mining framework is limited to analyzing occurrence counts of items in the Database. However, this framework applies a single minimum utility threshold value that fails to consider different item characteristics. Recent methods of association mining focused on finding the high utility itemsets instead of frequent itemsets generations. Some utility-based mining methods that is Faster High Utility Itemset Mining (FHM), High Utility Itemset Miner (HUI-Miner), Direct Discovery of High Utility Patterns (D2HUP), Utility Pattern Growth (UP Growth & UP Growth+) are studied for the generation of high utility itemsets generations. Existing HUI mining methods are effectively generating HUIs. However, developing a faster and memory-efficient HUI mining method is required. For this purpose, this work develops an Enhanced Fast - High Utility Itemset Mining (EF-HUIM) method for the faster generation of high utility itemsets and respective association rules.
Keywords: Association Mining, Utility, HUI, Data Mining, Item set, Association Rules.

1. Introduction

Association rule learning [1] is divided into three types of algorithms as Apriori, Eclat, F-P Growth Algorithm. Apriori and FP-growth uses the support and confidence, in this process the items are not prioritized that is the items are considered with the same level of importance. These algorithms are used in mining frequent products sets and relevant association rules. Apriori algorithm works [2] on databases that contains transactions, it uses Breadth First Search and Hash Tree to calculate frequent itemsets. It is simple to understand and apply, mainly used in market basket analysis. Nevertheless, an expensive method. Eclat Algorithm that is Equivalence Class Transformation uses a Depth First Search technique to find frequent itemsets. It performs faster execution than Apriori Algorithm. Whereas F-P growth algorithm stands for Frequent Pattern and it is the improved version of the Apriori Algorithm. It represents the database in the form of a tree structure, when database is large it becomes difficult to build and expensive. And these are employed in Market Basket analysis [3], Web usage mining, continuous production, etc. With the help of ARM techniques, we can know which items are sold together and the main goal of these techniques is to extract all the frequent itemsets. Sometimes, only the frequency is not a major concern in the market analysis. Profit or Utility analysis is also required for improving the market strategies. High Utility Itemset [4] mining is considered to be one of the important and challenging problems in data mining. The existing mining framework is limited to the analysis of occurrence counts of items in the database. However, this framework applies a single minimum utility threshold value that fails to consider different item characteristics. Here the problem offers greater flexibility to a decision-maker in using item utilities such as profits and margins to mine interesting and actionable patterns from databases. High utility itemset (HUI) mining is one of the important problems in data mining. A high utility itemset (HUI) is an itemset whose utility value is above a user-specified utility threshold. It can be considered as a generalized version of the frequent itemset. An itemset mining framework that is based on HUI offers greater flexibility to a decision-maker in incorporating her/his notion of item utilities such as margins, profits, and so on. The traditional frequent itemset mining framework is limited to the analysis of occurrence counts of items in the database. Recent methods of association mining focused on finding the high utility itemsets instead of frequent itemsets generations. Some utility-based mining methods, that is Faster High Utility Itemset Mining (FHM) [5], High Utility Itemset Miner (HUI-Miner) [6], Direct Discovery of High Utility Patterns (D2HUP) [7], Utility Pattern Growth (UP Growth & UP Growth+) [8] are studied for the generation of high utility itemsets generations.

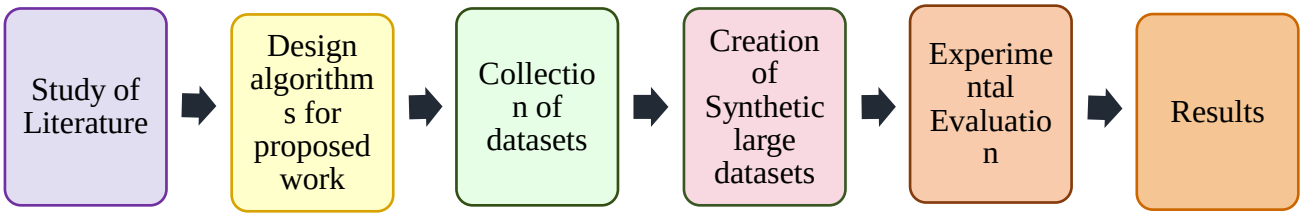
2. Related Work

UP-Growth (Utility Pattern Growth) [8], is the algorithm used for discovering high utility itemsets. Correspondingly, a compact tree structure called UP-Tree (Utility Pattern Tree), is used to maintain the important information of the transaction database related to the utility patterns. High utility itemsets can be generated from UP-Tree [9] by scanning the database. The utility pattern growth algorithm with more compressed utility pattern tree but it is time consuming as it recursively process all conditional prefix trees for candidate generation. UP-Growth+ is an improved version of UP-Growth with two pruning strategies. Information of high utility itemset is maintained in a special data structure named utility pattern tree (UP-Tree) and the candidate itemsets are generated with one scans of the database. The estimated utilities of candidates are well reduced, by discarding the utilities of the items which are impossible to be high utility or not involved in the search space. This not only

decreases the estimated utilities of the potential high utility itemsets but also reduces the number of candidates. However, it is still time consuming and high computational cost. Direct Discovery of High Utility Pattern (D2HUP) is algorithm which directly discover HUIs without maintaining candidates, it scans the whole database in one phase and finds the itemset which are high utility patterns but it consumes more memory. High Utility Itemset Miner (HUI-Miner) uses a structure called utility-list which is used to store both the utility information about an itemset and heuristic information for pruning the search space. HUI-Miner is the first to introduce the concept of remaining utility and vertical data structure but the join operations between utility lists of (k+1)-itemsets and k-itemsets is time consuming. HUI-Miner algorithm, generates a smaller number of candidates than UP-Growth and UP-Growth+. Fast High Utility Miner (FHM) which extends the HUI-Miner algorithm. It is a depth-first search algorithm that depends on utility lists to calculate the utility of itemsets. It integrates a novel strategy named estimated utility co-occurrence Pruning (EUCP) to reduce the number of joins operations while mining high utility itemset using utility list data structure. FHM algorithm [10] is up to 6 times faster than HUI-Miner algorithm. However, it consumes slightly more memory than HUI-Miner as well as poor performance on dense datasets.

3. Proposed Work

Extraction of high utility itemsets (HUIs) is the key objective of the HUI mining methods. Existing HUI mining methods are effectively generating HUIs. Nevertheless, it is required to develop a faster and memory-efficient HUI mining method [11]. For this purpose, this project work develops an Enhanced Fast - High Utility Itemset Mining (EF-HUIM) method for the faster generation of high utility item sets in an efficient way. An Enhanced Fast - High Utility Itemset Mining (EF-HUIM) method for the faster generation of high utility itemsets and respective association rules. It relies on two new upper-bounds named sub-tree utility and local utility to prune the search space. It also introduces a novel array-based utility counting approach named Fast Utility Counting to calculate these upper-bounds in linear time and space. Moreover, to reduce the cost of database scans, EF-HUIM introduces two techniques for database projection and transaction merging named High-utility Database Projection (HDP) and High-utility Transaction Mergin (HTM), also performed in linear time and space. An extensive experimental study on various datasets shows that EF-HUIM is in general two to three orders of magnitude faster and consumes up to eight times less memory than the state-of-art algorithms UP-Growth, UP-Growth+, d 2HUP, HUI-Miner and FHM.



High utility itemset (HUI) mining is one of the important problems in data mining. Recent methods of association mining focused on finding the high utility itemsets instead of frequent itemsets generations. Some utility-based mining methods, that is Faster High Utility Itemset Mining (FHM), High Utility Itemset Miner (HUI-Miner), Direct Discovery of High Utility Patterns (D2HUP), Utility Pattern Growth (UP Growth & UP Growth+) are studied for the generation of high utility itemsets generations. However, it is required to develop a faster and memory-efficient HUI mining method. For this purpose, this project work develops an Enhanced Fast - High Utility Itemset Mining (EF-HUIM) method for the faster generation of high utility itemsets and respective association rules.

4. Datasets

Algorithms were implemented in Java, and memory measurements were done using the standard Java API. Experiments were performed using a set of standard datasets, namely (Accident, Chess, Foodmart and Mushroom). These datasets are chosen because they have varied characteristics. Tables 1 and 2 summarize their characteristics. Accident is a large dataset [12] with moderately long transactions. Chess is a dense dataset with long transactions and few items. Foodmart is a sparse dataset with short transactions. Mushroom is a dense dataset with moderately long transactions. Foodmart is a customer transaction database containing real external/internal utility values.

TABLE 1. Dataset characteristics [12]

Dataset	#Transactions	#Distinct items	Avg. trans. length
Foodmart	4,141	1,559	4.4
Mushroom	8,124	119	23.0
Chess	3,196	75	37.0
Accident	340,183	468	33.8

5. Experimental Results

We compare the time and space taken by these algorithms to find high utility item sets on the same datasets. Firstly, when we consider the time taken by these algorithms to find high utility item sets from the datasets, i.e., Food mart, Mushroom, Chess, and Accident. The following results are obtained.

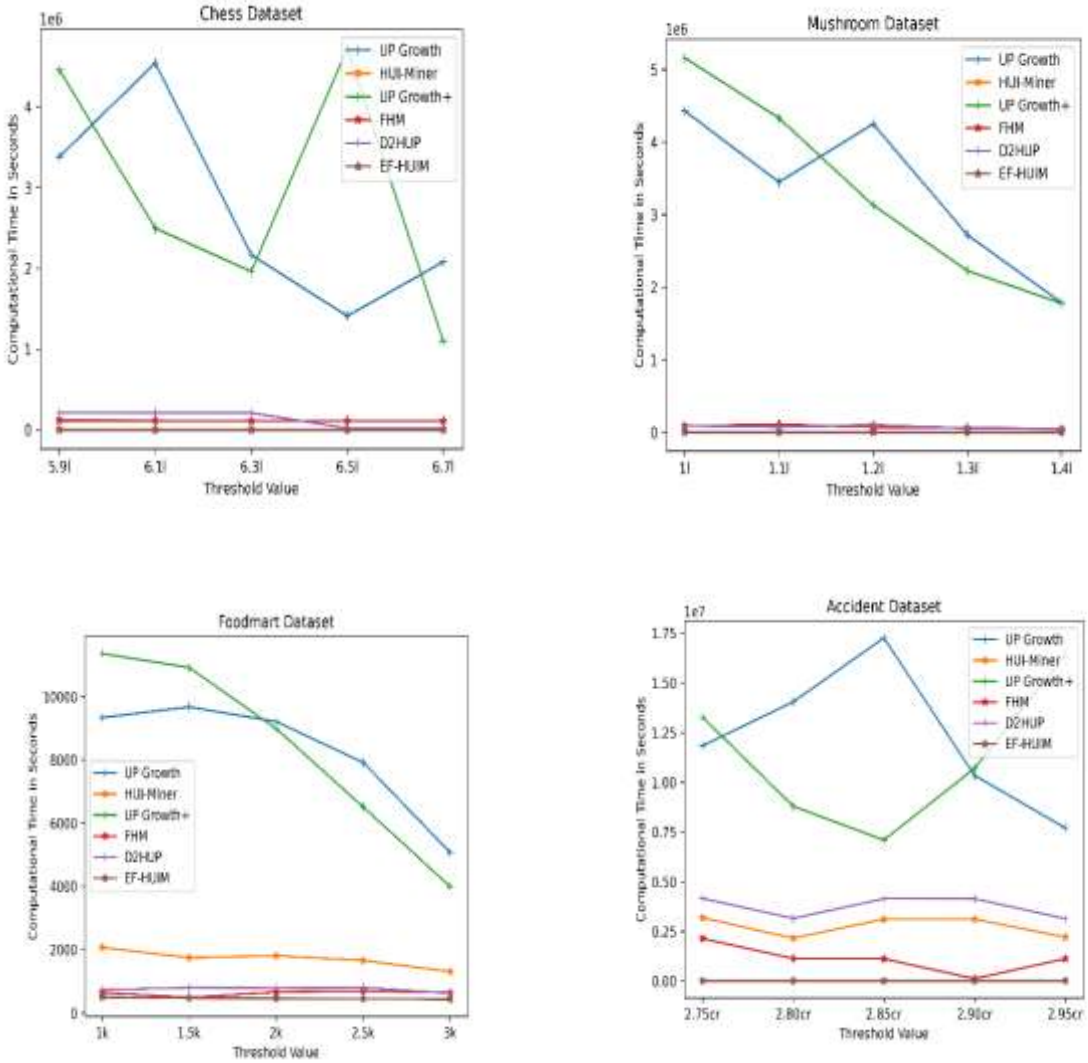


FIGURE 1. Execution time comparison on the datasets

We can observe that the time is taken by the Enhanced Fast - High Utility Itemset Mining (EF-HUIM) algorithm for generating high utility itemset is less compared to the remaining five algorithms; which traditional methods are as follows [13] [14]: Faster High Utility Itemset Mining (FHM), High Utility Itemset Miner (HUI-Miner), Direct Discovery of High Utility Patterns (D2HUP), Utility Pattern Growth (UP Growth & UP Growth+). The following results are obtained when we

consider the space taken by these algorithms to find high utility itemsets from the datasets, i.e., Foodmart, Mushroom, Chess, and Accident.

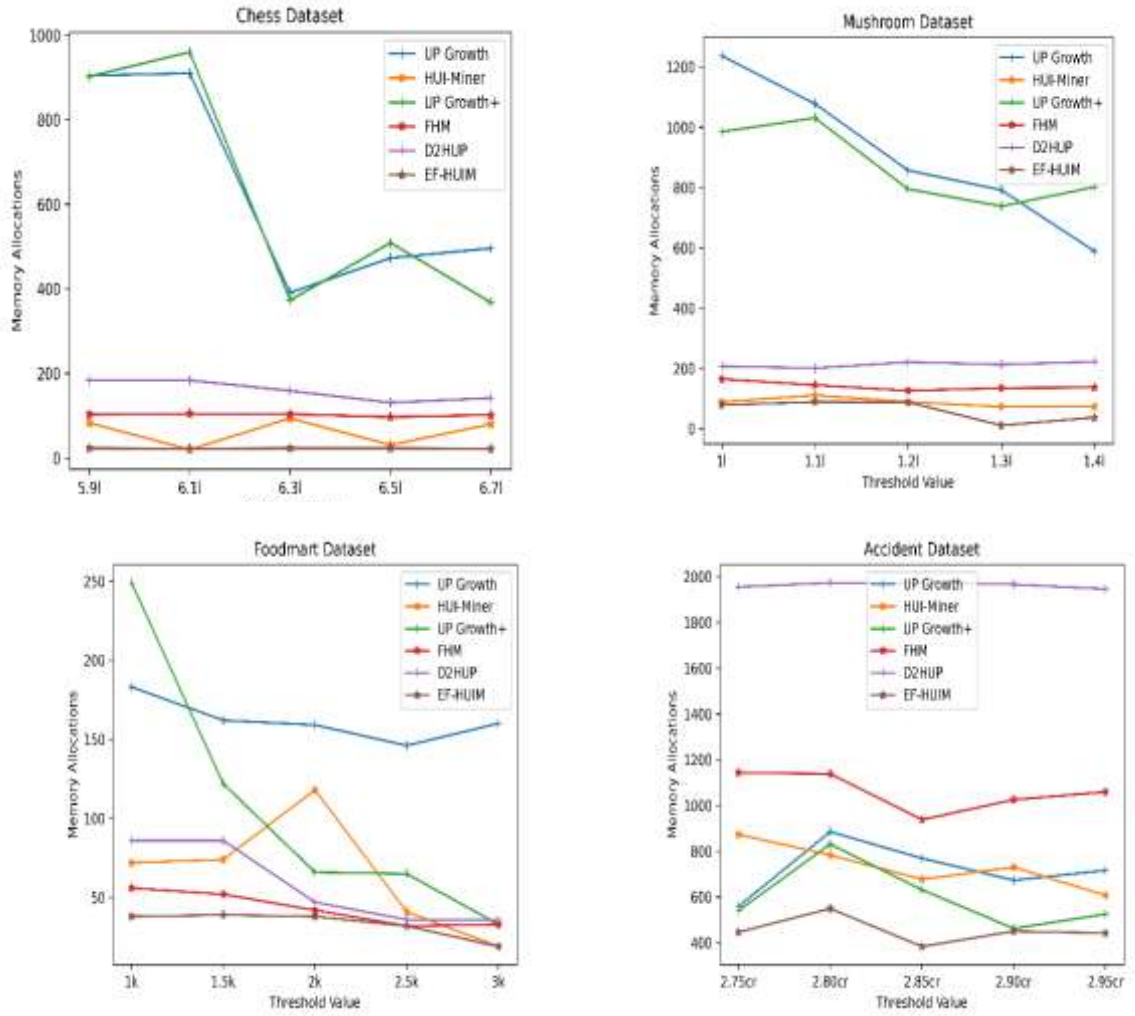


FIGURE 2. Memory Consumption comparison on the datasets

We can observe that the space is taken by the Enhanced Fast - High Utility Itemed Mining (EF-HUIM) algorithm for generating high utility item set is less compared to the remaining five algorithms; which existing algorithms are as follows: Faster High Utility Item set Mining (FHM), High Utility Itemed Miner (HUI-Miner), Direct Discovery of High Utility Patterns (D2HUP), Utility Pattern Growth (UP Growth & UP Growth+). From the above graphs, we can say that the Enhanced Fast - High Utility Item set Mining (EF-HUIM) algorithm gives impressive results, i.e., it is efficient in terms of time and space. We can observe that the EF-HUIM algorithm generates high utility item sets in less time and less space compared to the remaining five algorithms; which algorithms are as follows: Faster High Utility Itemed Mining (FHM), High Utility Item set Miner (HUI-Miner), Direct Discovery of High Utility Patterns (D2HUP), Utility Pattern Growth (UP Growth & UP Growth+).

6. Conclusion

Mining high utility item sets from transactional databases is performed to the purpose of discovery of item sets (with high utility-like profits). Although several relevant algorithms have been proposed in recent years, they incur the problem of producing many candidate item sets. Such a large number of candidate's item sets degrades the mining performance in terms of execution time and space requirements. The situation may worsen when the database contains lots of long transactions or long high utility item sets. For this purpose, our work develops an Enhanced Fast - High Utility Item set Mining (EF-HUIM) method for the faster generation of high utility item sets and generation of respective association rules. It relies on two new upper-bounds named sub-tree utility and local utility to prune the search space. It also introduces a novel array-based utility counting approach, fast utility counting, to calculate these upper-bounds in linear time and space. Moreover, to reduce the cost of database scans, EF-HUIM introduces two techniques for database projection and transaction merging named High-utility database projection (HDP) and High-utility transaction margin (HTM), also performed in linear time and space. An extensive experimental study on various datasets is done and it proven that EF-HUIM is two to three times faster and

consumes up to eight times less memory than the state-of-art algorithms UP-Growth, UP-Growth+, d2HUP, HUI-Miner, and FHM.

References

[1]. aki, M.J., Parthasarathy, S., Ogihara, M., Li, W. 1997. "Parallel algorithm for discovery of association rules." International Journal of Data mining and Knowledge Discovery, Vol.1, No.4, pp.343-374.

[2]. Zaki, M.J. 2001. "SPADE: An Efficient Algorithm for Mining Frequent Sequences." Machine Learning Journal, Vol.42, No.1-2, pp.31-60

[3]. Agarwal, R., Imielinski, T., Swami,A,N. 1993. "Mining association rules between sets of items in large databases." Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, Washington, D.C., United States, May 26-28, pp.207-216.

[4]. Narasimhulu K et al (2021) An enhanced cosine-based visual technique for the robust tweets data clustering. Int J IntellComputCybern 14(2):170–184

[5]. Deepa, N., Asmat Parveen, Anjum Khurshid, M. Ramachandran, C. Sathiyaraj, and C. Vimala. "A study on issues and preventive measures taken to control Covid-19." In AIP Conference Proceedings, vol. 2393, no. 1, p. 020226. AIP Publishing LLC, 2022.

[6]. Basha, M.S., Mouleeswaran, S.K. & Prasad, K.R. Detection of pre-cluster nano-tendency through multi-viewpoints cosine-based similarity approach. Nanotechnol. Environ. Eng. 7, 259–268 (2022). https://doi.org/10.1007/s41204-022-00222-8

[7]. Baralis, E. and Psaila, G. 1997. "Designing Templates for Mining Association Rules." Journal of Intelligent Information Systems, Vol.9, Issue.1, pp.7-32

[8]. M.P. Jenarthanan, N G Ramkhi, M. Ramachandran, Vimala Saravanan, "Mechanical, Morphological and Water absorption properties of Polypropylene based Composites", Materials and its Characterization, 1(1), (2022):48-52

[9]. Prasad K, Mohammed M, Prasad L, Anguraj DK (2021) an efficient sampling-based visualization technique for big data clustering with crisp partitions. Distrib Parallel Databases. https://doi.org/10.1007/s10619-021-07324-3

[10]. Borgelt, C., Kruse, R. 2002. "Induction of association rules: Apriori implementation." Proceedings of the Fifteenth Conference on Computational Statistics, Berlin, Germany, August 24-28, pp.395–400

[11]. Cheung, D.W., Han, J., Ng, V.T., Fu, A.W., Fu, Y. 1996. "A fast distributed algorithm for mining association rules." Proceedings of the Fourth International Conference on Parallel and Distributed Information Systems, Miami Beach, Florida, United States, December 18 - 20, pp.31-43.

[12]. R. Dhaneesh, Iswarya V.S, D.R. Pallavi, Ramachandran, Vimala Saravanan, "The Impact of Self-help Groups on the Women Empowerment in Tamil Nadu", Trends in Banking, Accounting and Business, 1(1), (2022):1-5

[13]. Basha, M.S., Mouleeswaran, S.K. & Prasad, K.R. Sampling-based visual assessment computing techniques for an efficient social data clustering. J Supercomput 77, 8013–8037 (2021). https://doi.org/10.1007/s11227-021-03618-6

[14]. Krishna, M.H., Dasore, A., Rajak, U., Konijeti, R., Verma, T.N. (2022). Thermo-Economic Optimization of Spiral Plate HX by Means of Gradient and Gradient-Free Algorithm. In: Verma, P., Samuel, O.D., Verma, T.N., Dwivedi, G. (eds) Advancement in Materials, Manufacturing and Energy Engineering, Vol. II. Lecture Notes in Mechanical Engineering. Springer, Singapore. https://doi.org/10.1007/978-981-16-8341-1_48

[15]. N. subash, M. Ramachandran, Vimala Saravanan, Vidhya prasanth, "An Investigation on Tabu Search Algorithms Optimization", Electrical and Automation Engineering, 1(1), (2022):13-20

[16]. Padmanabhan, B., Tuzhilin, A. 1998. "A belief-driven method for discovering unexpected patterns." Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining, New York city, USA, August 27-31, pp.94-100.

[17]. Suleman Basha, M., S. K. Mouleeswaran, and K. Rajendra Prasad. "Cluster tendency methods for visualizing the data partitions." Int J Innovative Technol Exploring Eng 8.11 (2019): 2978-2982.

[18]. Song, M., Rajasekaran, S. 2005. "Finding frequent itemsetsby transaction mapping." Proceedings of the twentieth ACM Symposium on applied computing, Santa Fe, New Mexico, March 13-17, pp.488-492.

