Mining High Utility Sequential Pattern using Lexicographic q-Sequence Tree and Utility Linked-List

Sreedevi Devaraj
Department of Computer Science and Engineering
Mangalam College of Engineering, Ettumanoor
Kottayam, India

Vivek M. K
Department of Computer Science and Engineering
Mangalam College of Engineering, Ettumanoor
Kottayam, India

Reno Rajan
Department of Computer Science and Engineering
Mangalam College of Engineering, Ettumanoor
Kottayam, India

Dr. Ranju S. Kartha
Department of Computer Science and Engineering
Mangalam College of Engineering, Ettumanoor
Kottayam, India.

***Abstract*—High utility sequential pattern (HUSP) mining is an important field to discover high utility patterns in a sequence. Nowadays it becomes more relevant and important in real life applications like market basket analysis, e-commerce recommendations and bio informatics etc. Sequential Pattern Mining (SPM) is used to mine or extract sequential or frequent patterns from vast database. In traditional SPM certain factors like utility of products, profit are not considered. To improve its features, the process of SPM is generalized to HUSP Mining (HUSPM) which is used to discover the high utility patterns in a sequence database. Many algorithms have been proposed to find the high utility of a sequence database, but due to the large search space, the combinatorial explosion has been raised. This paper proposes a new algorithm, for mining HUSP-Utility Linked List (ULL). The objective of HUSP-ULL is to discover the sequential pattern and to find the utility of each pattern in the database, that meets or exceeds predefined minimum utility threshold. HUSPM make use of lexicographic q-sequence and UL (Utility Linked) - list for identifying high utility patterns. The obtained output can be used in many applications like e-commerce, market basket analysis, healthcare industry, web** mining**, bioinformatics and mobile computing etc.**

Keywords— Lexicographic q- Sequence, Utility Linked List structure, High Utility Pattern Mining, Pruning Strategy, Sequential Pattern Mining.

I. INTRODUCTION

Sequential pattern mining (SPM) [1],[7],[8] is an emerging and interesting area of research in extracting the knowledge or information in a database. Utility mining is a new approach in data mining where mining results must meet user's goals. Existing algorithms of association rule, mining does not consider interestingness measures for users. Previously many algorithms were proposed for frequent pattern mining, but most of them mainly based on the count or occurrence value of an itemset. In this paper, a new approach for high utility pattern mining has been proposed which uses pruning and bagging methods to improve the performance.

Utility based pattern mining in sequential database is more challenging than frequent itemset mining [1], [9] and ordinary SPM. Consider two products TV and milk bottle, when mining the database milk bottle may appear more frequent than TV but the utility of TV is more than that of milk bottles. So, in such situations mining became more complicated. High utility sequential pattern mining (HUSPM) [4][6][9][10] generalizes SPM and it is used for mining sequence pattern by considering the utility. So. HUSPM extracts high utility sequential patterns that can be used in many real time applications.

In market basket analysis it can easily identify the association between two products or pair of products purchased and also identify pattern of co-occurrence. It means that two or more process occurring parallelly. Consider an item X is purchased by customer and an item Y is likely to be purchased and that it will be based on the probability rules derived from frequency of co-occurrence. So, by applying various cross selling strategies, the selling such products can be improved. In e commerce recommendation system, it helps the customer to find and purchase product using e-commerce sites. HUSP mining helps them to discover the most relevant search results and also helps to promote the products. So, this can be turned into a serious business tool. Consider an example- a book recommendation in a library. The student is spending a lot of time for searching a particular book in a library and sometimes it may not be available or it is difficult to find. The algorithm extracts the frequent pattern of books that may help the students to save their valuable time.

The main objective of the proposed algorithm is to:

- Discover the sequential pattern and find the utility of each pattern in the database, that meets or exceeds predefined minimum utility threshold.
- Reduce the time and memory required when compared to other algorithms.

In this paper, the remaining sections are organized as follows. Section II comprises of the literature review of related works. Research Methodologies are presented in Section III. Result Analysis based on the proposed algorithm is provided in Section IV. Finally, conclusions are drawn in Section V.

II. LITERATURE SURVEY

Many algorithms are available to find the high utility of a sequence database. But finding the high utility pattern in a sequence database is a complex task. Efficiency and scalability are the major insight for all the algorithms. Earlier, the algorithms of HUSPM like UP and UL [11] which performs breadth-first-search and depth-first-search respectively. These algorithms discover patterns and high utility in two different phases. Also, these two algorithms are very time consuming and wastage of more memory spaces. Other algorithms like USpan [12], HUSP uses SWU (sequence-weighted utilization) and it generates too many candidates so the efficiency is very less. The major disadvantage of USpan is that it is not a complete algorithm for generating HUSP.

PrefixSpan [2] mines the complete set of patterns and it reduces the efforts of a person's subsequence generation. Prefix-projection substantially reduces the size of projected databases and leads to efficient processing. But there are some time related issues. Frequent pattern tree (FP-tree) [1] structure, which is an extended prefix tree structure for storing compressed and, crucial information about frequent patterns. Here, the number of combinations to be found is comparatively less. The concept of UP_Growth (Utility Pattern Growth) [3] for mining high utility item sets were proposed. But the accuracy related issues are raised in this algorithm. Later an algorithm to mine Top-k high utility itemset [4] were proposed by Tseng V. S. et.al. It consists of two algorithms named TKU (mining Top-K Utility item sets) and TKO (mining Top-K utility item sets in One phase) for mining. Unfortunately, it is performing mining process of k-top high utility itemset only. A two-phase algorithm MHUH [5] were proposed by P. Fournier-Viger et. al. The first phase named Extension, the existing algorithm FHUSpan [5] efficiently mine the general high-utility sequences (g-sequences). The second phase named Replacement; the special high-utility sequences is mined with the hierarchical relation (s-sequences) as high-utility hierarchical sequential patterns from g-sequences. Here the accuracy related issues are the major challenge.

Projection-based Utility (ProUM) is an approach to find high-utility sequential pattern from a sequence of data [15]. The limitation of this approach is when dealing with sequence data since they are time-consuming and require large amount of memory usage. An algorithm named fast algorithm for mining discriminative high utility patterns (DHUPs) with strong frequency affinity (FDHUP) [16] is proposed to efficiently discover DHUPs by considering both the utility and frequency affinity constraints. Two compact structures named EI-table and FU-tree and three pruning strategies are introduced in the proposed algorithm to reduce the search space, to discover DHUPs. But, it is not as much efficient as the algorithm proposed in this article. U. Yun, D. Kim, E. Yoon, and H. Fujita introduced a method called high average utility pattern mining (HAUPM)[18] approach, which discovers patterns that are related to one another. Eventhough, this method provides important patterns, the search space cannot be reduced. An algorithm proposed by W. Gan called High Utility Occupancy Pattern Mining (HUOPM)[19] uses two data structures called utility occupancy list and frequency utility table for effectively finding useful patterns without candidate generation. But the issue is that it can be only used in static databases. An algorithm called High Incremental High Utility Itemset Mining (iHUIM) [20] were proposed which incrementally updates and outputs the high utility itemsets and a dynamic dataset is used rather than a static dataset. But there is no downward closure property to reduce the search space.

In the case of top-k high utility itemset mining, where k is the desired number of high utility item sets to be mined. The proposed algorithms have good scalability on large datasets and the performance of the proposed algorithms is very close to the optimal case of the state-of-the-art of first and second phases of the existing utility mining algorithms. Although here it is proposing a new framework for top-k HUI [4] mining, it has not yet been incorporated with other utility mining tasks to discover different types of top-k high utility patterns such as top-k high utility episodes, top-k closed high utility itemset, top-k high utility web access patterns and top-k mobile high utility sequential patterns. These gives an opportunity for exploration as its future work.

III. SIGNIFICANCE, CONTRIBUTIONS AND METHODOLOGY

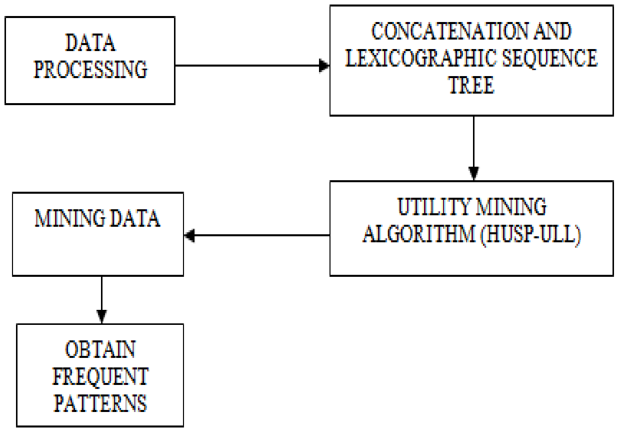


Fig. 1. Proposed System Architecture

Fig.1 shows the system architecture of the proposed system, it consists of following modules:

A. Concatenation and Creation of LQS- tree

LQS- tree is mostly used in most of the HUSPM algorithms to represent the search space [12] for HUSPs. In this case, each node is used to represent a candidate of HUSP whose utility value will be compared with the minimum threshold value to check whether it is HUSP or not. To add a new node to the LQS- tree, two operations [12][13] are performed and they are I- Concatenation and S- Concatenation.

In I- Concatenation, the new item is appended along with the last item in the sequence. In S- Concatenation, the new item is appended to the sequence as the last element.

Therefore, the number of items in I- Concatenation remains the same while the number of items in S- Concatenation grows. So, the result of these two operations will generate a sequence which is the search space for mining HUSPs.

B. UL-List

Utility Linked List [13] is used to record the information about the utility of each sequence that has been generated during the concatenation operations. These UL - list consist of two parts: the first one is the Header Table, which is used to store set of items with its first occurring place in transformed sequence and, the second one is the Utility and Position (UP) information, which stores the details about utility of certain item.

C. Closure Property of Upper Bound

This paper proposes a downward closure property called Sequence- Weighted Utilization (SWU) [12] upper bound to identify HUSP using the algorithm HUSP- ULL without any combinatorial explosion of search space. By creating an upper bound, each space can be reduced and it increase the speed of the mining process.

D. Pruning Itemset

A LAR Strategy [14] is used to remove candidate item from a sequence. So, only less umber of item will be considered while concatenating two operations. This strategy reduces the execution time required by the algorithm.

E. HUSP- ULL Algorithm

HUSP- ULL algorithm [14] is mainly used for scanning the sequential database and generates the UL- list for each of the q- sequences. Further, the algorithm works based on the above provided four factors. At last, the algorithm will provide the set of HUSPs that has been discovered as the output.

The algorithm firstly removes unwanted items and then recalculates the UL-list. Each node in a lexicographic q-sequence (LQS)- tree represents a candidate HUSP, whose utility can be compared with the minimum utility threshold to determine if the candidate is a HUSP. For each node that the algorithm visits in the LQS-tree, a projected database is built, which consists of utility-linked (UL)-lists obtained by transforming transactions (q-sequences) of the original database. The algorithm utilizes the UL-lists of each node (candidate HUSP) present in the tree to calculate its utility and upper-bounds. Each UL-list represents a transaction (*q*-sequence). To add a new node to the LQS- tree, two operations are performed and are called as I- Concatenation and S- Concatenation. respectively. Then by using the LAR strategy, items having utility value less than minimum utility value are discarded. After, that using a Judge procedure, the candidates having utility values not less than minimum threshold value are generated and provided as the output.

Based on the LQS-tree and UL-lists, the proposed HUSP-ULL algorithm can successfully identify the complete set of HUSPs using a depth-first search that applies two concatenation operations. However, this process can lead to exploring a very large number of candidates in the LQS-tree, since there is a combinational explosion of the number of candidates in the mining process of HUSPs. To speed up the mining process and also to maintain the downward closure property, a new property called sequence-weighted utilization (SWU) upper-bound was proposed to obtain a sequence-weighted downward closure (SWDC) property for HUSPs mining. This property greatly helps in reducing the search space and eliminate unpromising candidates early for reducing execution time.

IV. Result and Discussions

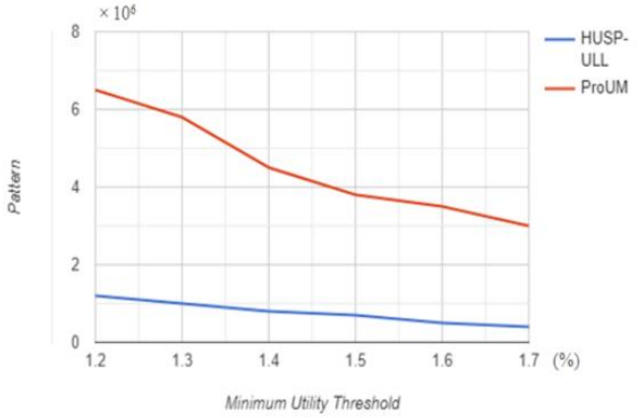
The proposed algorithm is used to obtain high utility sequential patterns (HUSPs). Here, three datasets were used for performing the experiments. Among the three datasets, all of them are real-life datasets. The datasets that are used are as shown in the following TABLE I:

TABLE I. DATA SET DETAILS

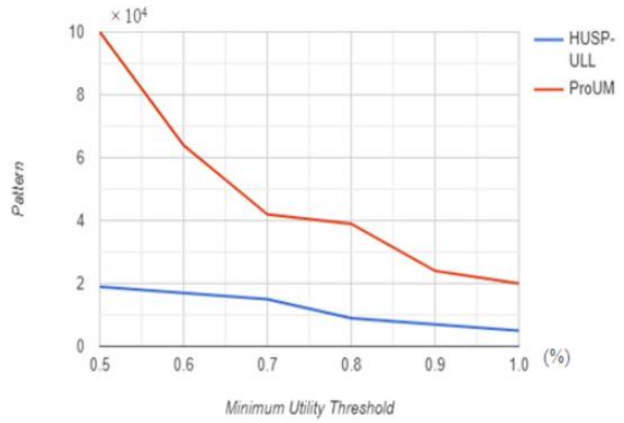
Dataset	***No. of Sequence***	***Avg. no. of elements per sequence***
Sign	730	52.0
Bible	36,369	21.6
Leviathan	5834	33.8

- Sign:It is a real-life dataset which contains sequence of sign language utterance.
- Bible:It is a real-life dataset that is prepared by converting the bible into a set of sequence of words.
- Leviathan:It is a conversion of one of the work of Thomas Hobbes' Leviathan to a sequence of words.

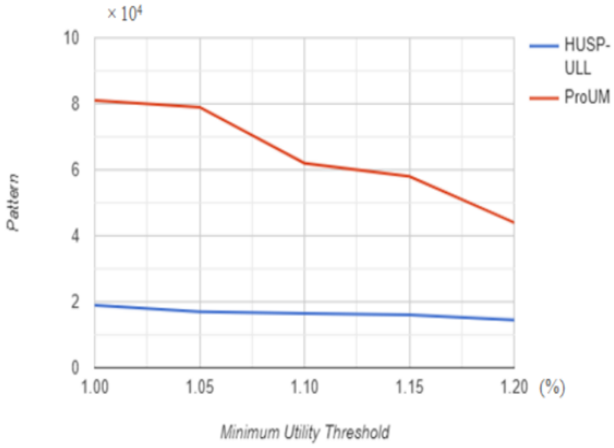
The *Fig. 2* represents the pattern mined details from three datasets sign(*Fig.2(a)*), bible(*Fig.2(b)*) and Levithan(*Fig.2(c)*) by using two algorithms, HUSP-ULL and ProUM.



(a)



(b)



(c)

Fig. 2. Comparative Study on different Dataset

The HUSPs obtained from the proposed algorithm can be applied in various real-life applications that are already mentioned in this paper. Here the transaction done by the user is applied as the input to the proposed algorithm. The algorithm calculates the itemset utility value. The frequently made transactions can be find out from the output.

Suppose, Products Purchased in a supermarket:

Transaction 1:
- **C**oke, 6
- Chips, 2
- Dip, 1

Transaction 2:
- Coke, 1

Transaction 3:
- Coke, 2
- Chips, 1

Transaction 4:
- Chips, 1

Transaction 5:
- Chips, 2

Transaction 6:
- Coke, 6
- Chips, 1

After applying the proposed algorithm HUSP- ULL in this dataset the output shown as below:

Frequent Items=('Chips', 'Coke'), itemset_utility=30.02
Frequent Items=('Chips',), itemset_utility=20.93

By using the pruning strategy in this algorithm, the number of candidates that are generated in these three datasets are comparatively less than the number of candidates generated by other existing algorithms. So, from this also we can clearly say find that the memory usage of the proposed algorithm is far less than the existing algorithms. Therefore, if is identified that this algorithm is useful for pattern mining in large datasets.

V. CONCLUSION

The utility-based sequence pattern mining is a vital issue seen among certain real-life applications. The algorithm HUSP- ULL, proposed in this article provide an efficient output as desired. Experiments done on different datasets using the proposed algorithm showed the efficient and effective identification and retrieval of HUSPs. Also, this algorithm reduces the search space by the use of pruning strategy. The only concern that arises in this scenario is the accuracy variation depends on the datasets that are available. This article also provides certain ideas that can be used for future developments in this sequential pattern mining field.

VI. FUTURE SCOPE

The proposed algorithm in this paper is a pattern mining algorithm and here it can be used for classification purpose. Here it is used to recognize the patterns and chooses the frequently made transactions. So, in the future it can be modified by including performance matrix and comparative study of various algorithms. So the data can be divided as training and testing sets and by using the performance matrix and resultant graphs the classification of data can be done more accurately and efficiently.

ACKNOWLEDGEMENT

We would like to express our special thanks of gratitude to Dr. Vinodh P. Vijayan for his able guidance and support in our research work. We would also like to extend our gratitude to the Principal Dr. Manoj George (Mangalam College of Engineering, Ettumanoor, Kerala) for providing us with all the facilities that are required.

REFERENCES

[1] J. Han, J. Pei, Y. Yin, and R. Mao, "Mining frequent patterns without candidate generation: A frequent-pattern tree approach," Data Min. Knowl. Disc., vol. 8, no. 1, pp. 53–87, 2004.

[2] J. Pei et al., "PrefixSpan: Mining sequential patterns efficiently by prefix projected pattern growth", Proc. IEEE Int. Conf. Data Eng.1, pp. 215–224, 2001.

[3] J.K.kavithaa, D.Manjulab and J.K.Kasthuribha "Fast Update Utility Pattern Tree (FUUP – Tree)", International Conference on Mathematical Computer Engineering - ICMCE – 2013.

[4] V. S. Tseng, C.-W. Wu, P. Fournier-Viger, and P.-S. Yu, "Efficient algorithms for mining top-k high utility itemsets," IEEE Trans. Knowl. Data Eng., vol. 28, no. 1, pp. 54–67, Mar. 2017.

[5] P. Fournier-Viger, J. C.-W. Lin, R.-U. Kiran, and Y.-S. Koh, "A survey of sequential pattern mining," Data Sci. Pattern Recognit., vol. 1, no. 1,pp. 54–77, 2017.

[6] Chunkai Zhang , Zilin Du , and Yiwen Zu "An Efficient Algorithm for Extracting High-Utility Hierarchical Sequential Patterns",Volume 2020, Article ID 8816228 July 2020.

[7] R. Agrawal and R. Srikant, "Mining sequential patterns,"in *Proc. IEEE Int. Conf. Data Eng.*, 1995, pp. 3–14.

[8] R. Srikant and R. Agrawal, "Mining sequential patterns:Generalizations and performance improvements," in *Proc. Int. Conf. Extend. Database Technol.*, 1996, pp. 1–17.

[9] R. Chan, Q. Yang, and Y.-D. Shen, "Mining high utility itemsets," in *Proc. 3rd IEEE Int. Conf. Data Min.*, 2003, pp. 19–26.

[10] G.-C. Lan, T.-P. Hong, V.-S. Tseng, and S.-L. Wang, "Applying the maximum utility measure in high utility sequential pattern mining," *Expert Syst. Appl.*, vol. 41, no. 11, pp. 5071–5081, 2014.

[11] C. F. Ahmed, S. K. Tanbeer, and B. S. Jeong, "A novel approach for mining high-utility sequential patterns in sequence databases," *ETRI J.*, vol. 32, no. 5, pp. 676–686, 2010.

[12] J. Yin, Z. Zheng, and L. Cao, "USpan: An effificient algorithm for mining high utility sequential patterns," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Disc. Data Min.*, 2012, pp. 660–668.

[13] J.-Z. Wang, J.-L. Huang, and Y.-C. Chen, "On efficiently mining high utility sequential patterns," *Knowl. Inf. Syst.*, vol. 49, no. 2, pp. 597–627, 2016.

[14] Wensheng Gan, Jerry Chun-Wei Lin, Jiexiong Zhang, Philippe Fournier-Viger , Han-Chieh Chao and Philip S. Yu "Fast Utility

Mining on Sequence Data", IEEE TRANSACTIONS ON CYBERNETICS, 2020

[15] W. Gan, J. C.-W. Lin, J. Zhang, H.-C. Chao, H. Fujita, and P.-S. Yu, "ProUM: High utility sequential pattern mining," in *Proc. IEEE Int. Conf. Syst. Man Cybern.*, 2019, pp. 767–773.

[16] J. C.-W. Lin, W. Gan, P. Fournier-Viger, T.-P. Hong, and H.-C. Chao, "FDHUP: Fast algorithm for mining discriminative high utility patterns," *Knowl. Inf. Syst.*, vol. 51, no. 3, pp. 873–909, 2017.

[17] W. Gan, J. C.-W. Lin, P. Fournier-Viger, H.-C. Chao, and H. Fujita, "Extracting non-redundant correlated purchase behaviors by utility measure," *Knowl. Based Syst.*, vol. 143, pp. 30–41, Mar. 2018.

[18] U. Yun, D. Kim, E. Yoon, and H. Fujita, "Damped window based high average utility pattern mining over data streams," *Knowl. Based Syst.*, vol. 144, pp. 188–205, Mar. 2018.

[19] W. Gan, J. C.-W. Lin, P. Fournier-Viger, H.-C. Chao, and P. S. Yu, "HUOPM: High-utility occupancy pattern mining," *IEEE Trans. Cybern.*, vol. 50, no. 3, pp. 1195–1208, 2020, doi: 10.1109/TCYB.2019.2896267.

[20] W. Gan, J. C.-W. Lin, P. Fournier-Viger, H.-C. Chao, T.-P. Hong, and H. Fujita, "A survey of incremental high-utility itemset mining," *Interdiscipl. Rev. Data Min. Knowl. Disc.*, vol. 8, no. 2, 2018, Art. no. e1242.