Implementation and Analysis of High Utility Sequential Pattern Mining Algorithms

Report submitted in fulfillment of the requirements
for the B.Tech. Project (7th semester)
by

Jatin Garg
Lakshay Jindal
Ojus Bhutani

under the guidance of

Dr. Bhaskar Biswas

Department of Computer Science and Engineering
INDIAN INSTITUTE OF TECHNOLOGY (BHU) VARANASI
Varanasi 221005, India

25 November 2022

Declaration

We certify that

1. The work contained in this report is original and has been done ourselves under the general supervision of our supervisor.
2. The work has not been submitted for any project.
3. Whenever we have used materials (data, theoretical analysis, results) from other sources, we have given due credit to them by citing them in the text of the thesis and giving their details in the references.
4. Whenever we have quoted written materials from other sources, we have put them under quotation marks and given due credit to the sources by citing them and giving required details in the references.

Jatin Garg
Lakshay Jindal
Ojus Bhutani
B.Tech Part(IV) Student
Department of Computer Science and Engineering
Indian Institute of Technology (BHU) Varanasi
Varanasi, INDIA 221005

Date: 25 November, 2022
Place: IIT (BHU) Varanasi

Certificate

This is to certify that the work contained in this report entitled ***Implementation and Analysis of High Utility Sequential Pattern Mining Algorithms (HUSPM)*** being submitted by ***Lakshay Jindal (19075042)***, ***Jatin Garg (19075037)*** and ***Ojus Bhutani (19075051)***, carried out in the Department of Computer Science and Engineering, Indian Institute of Technology (BHU) Varanasi, is a bonafide work of our supervision.

Date: 25 November, 2022
Place: IIT (BHU) Varanasi

Dr. Bhaskar Biswas
Department of Computer Science and Engineering
Indian Institute of Technology (BHU) Varanasi
Varanasi, INDIA 221005

Acknowledgements

We would like to express our sincere gratitude to our supervisor Dr. Bhaskar Biswas for his guidance and support during the timeline of the whole project.

Jatin Garg
Lakshay Jindal
Ojus Bhutani
B.Tech Part(IV) Student
Department of Computer Science and Engineering
Indian Institute of Technology (BHU) Varanasi
Varanasi, INDIA 221005

Date: 25 November, 2022
Place: IIT (BHU) Varanasi

Abstract

High utility sequential pattern mining is a data mining task similar to sequential pattern mining where we take an additional factor called utility value into account besides the frequency of items to generate sub-sequences. We need to find the sequences that have a total utility value above a threshold utility value called "minutility". The main task of High Utility Sequential Pattern Mining (HUSPM) is to find all sub-sequences, within a quantitative sequential database, that have high utility with respect to the minutility threshold. The output contains all the possible sub-sequences having utility value above this particular threshold. It has numerous applications in various fields like retail marketing and web service. It is very useful in business decision making where we have to take other factors like profit, cost, weight etc into account besides frequency of items. It also finds applications in bio-informatics, consumer behavior analysis, and web-page click-stream mining. Variety of algorithms have been proposed for high utility sequential pattern mining in the past decades. This report talks about another such algorithm: Projection-based Utility Mining on sequence data (ProUM)[1] .To improve efficiency, TKUS adopts a projection and local search mechanism and employs several schemes, including the Sequence Utility Raising, Terminate Descendants Early, and Eliminate Unpromising Items strategies, which allow it to greatly reduce the search space. We will analyze the efficiency of this algorithm and suggest a soft computing approach from results and observations.

Contents

1 Introduction

1.1 Overview

Data Mining is used to extract useful information and patterns from an existing database. It is an area of extensive research due to its applications in several domains like machine learning, text mining, medicine, surveillance and business. There are several data mining algorithms already present to analyze data. Their working vary depending upon the specific requirements of the task and size of database.

Frequent Pattern Mining is one of the most popular field of data mining. Frequent Pattern Mining algorithms analyze the data and try to find pattern and/or relationships between the data items based on different parameters. It can be of several types depending upon the types of patterns like Frequent Itemset Mining, Sequential Pattern Mining and Graph Mining.

Frequent itemset mining[2] is used to discover frequently appearing set of items in a transaction database. Transaction database consists of a set of items in each row indicating the items purchased by a customer in a particular transaction. It has variety of uses like to find popular items which are purchased frequently or to know customer demands for building a business model.

High-Utility-Itemset mining generalizes the concept of Frequent-Itemset-Mining by addition an additional parameter called utility value to generate itemsets. Some itemsets found in frequent itemset mining may be useless due to various restrictions caused in real life applications. High Utility Itemset mining is useful in such cases as it incorporates utility value of each object which may denote its importance, weight etc.

But HUIM can't deal with sequential databases where each item is associated with a timestamp. To deal with this problem, the concepts of Sequential Pattern Mining (SPM) and High Utility Sequential Pattern Mining (HUSPM) were introduced.

Sequential Pattern Mining(SPM)[3] is a type of Frequent Pattern

Mining. The purpose of sequential pattern mining is to discover all frequent sequences as sequential patterns, which reflect the potential connections within items, from a sequence database under the given minimum support[4][5]. An example of such a sequential pattern is that customers typically buy a phone, then a phone shell, and then a phone charger. Customers who buy some other commodities in between also support this sequential pattern. It has variety of uses like to find series of related items which are purchased frequently or to know customer demands for building a business model.

High-Utility-Sequential-Pattern-Mining (HUSPM) generalizes the concept of Sequential-Pattern-Mining (SPM) by addition of utility value to generate itemsets. It takes the frequency of items, their utility (can be weight, importance etc.) as well as the relationship between them into account while searching for potential sub-sequences of items. It incorporates the idea of HUIM and SPM, and can be thought of as an advanced version of HUIM.

Consider the following example. We have to transport necessary medicines to a nearby city. But due to COVID-19 restrictions, we can visit the city only once. Due to loading limit on the transport vehicle, we can only transport limited amount of supplies. So we need to select the medicines which can satisfy the requirements of the people to the highest possible extent and minimize the casualties.

A possible solution may be to assign a utility value to each medicine. The utility value will denote the seriousness of the disease the medicine is prescribed for. We will set a minutility threshold and use high utility itemset mining to generate itemsets. This way, we can find the medicines which are purchased in large amounts as well as those medicines which are required for critically ill patients.

But the problem with the above solution is that it doesn't take the relationship between medicines into account. There can be some medicines which can't be taken individually and require some additional medicines along with them. The doctor may need to prescribe a group of medicines frequently instead of just a single medicine. So, all of the three factors below are important:

- Frequency - No. of people who require a particular medicine.

- Utility Value - Importance of medicine/ Seriousness of Disease for which it is required.
- Medicine Sequence/Relationship- Medicines which are required to be administered together.

So, we will use HUSPM concept to solve this problem. The final subsequences in result will include frequently used medicines with high utility value (important medicines) and are related to each other.

1.2 Research Motivation

There is a lot of research that has been done and is going on in the field of high utility sequential pattern mining. Several algorithms have already been proposed to efficiently generate high utility sequential patterns from different sequential databases. But there is lack of proper implementation of these algorithms to mine data-sets efficiently in case of real world problems. Algorithms proposed are only in theory or there is lack of proper implementation.

Last semester we implemented and analysed one such algorithm. Algorithm belongs to family of hard computing algorithms. This semester we proposed soft computing algorithm and methods to design more such variations.

So, we have the following goals in mind:

- We aim to design a good soft computing approach to efficiently solve the problem of high utility sequential pattern mining while maintaining the accuracy of the results.
- We aim to find more such variations of the proposed algorithm.

1.3 Organisation of Project

We will first discuss some important terms related to high utility sequential pattern mining. Then, we will briefly define the work done in last semester and new leads made in this semester.

2 Problem Statement and Definitions

- **Definition 1** Let $I = \{i_1, i_2...i_n\}$ be the finite set of items. Then $X \subseteq I$ is called an *itemset*. A sequence s is an ordered list of item sets (also called elements or events). A group of the sequences stored with their identifiers (sid) is called a sequence database, denoted as D. Thus, a sequence database $D = \{s_1, s_2, \dots s_m\}$ is a set of sequences/tuples $(sid, e_1 \dots, e_j)$, where sid is a sequence id, and e_j is an element that contains a set of items belonging to I.

- **Definition 2** For the addressed problem of HUSPM, the processed database is the quantitative sequence database (or called q-database) that each item $i_j \in I(1 \leq j \leq n)$ in an element/itemset v is associated with a quantity (also called internal utility), denoted as $q(i_j, v)$. For convenience, we use "q-" to refer to the object associated with quantity throughout this paper. Thus, the term "q-sequence" means a sequence with quantities, and "sequence" means a sequence without quantities. Similarly, the "q-itemset" means an itemset having quantities, while an "itemset" does not have quantities.

- **Definition 3** (utility mining on sequence) A k-itemset, also called $k-q$-itemset, is an itemset that contains exactly k items. A k-sequence ($k-q$-sequence) is a sequence having k items. Let $q(i_j, v)$ be the quantity of (i_j) in a q-itemset v, and $pr(i_j)$ be the unit profit of i_j. The utility of a q-item i_j in a q-itemset v is denoted as $u(i_j, v)$, and defined as $u(i_j, v) = q(i_j, v)pr(i_j)$. The utility of a q-itemset v is denoted as $u(v)$ and defined as $u(v) = \sum_{i_j \in v} u(i_j, v)$. The utility of a q-sequence $s =< v_1, v_2, \dots, v_d >$ is $u(s) = \sum_{v \in s} u(v)$. The utility of a quantitative sequential database D is the sum of the utility of each q-sequences, such as $u(D) = \sum_{s \in D} u(s)$.

- **Definition 4** An itemset $X = \{x_1, x_2, \dots, x_m\}$ is said to be "contained" in q-itemset $I = \{i_1, i_2, \dots, i_n\}$ (i.e. $X \subseteq I$) if $n \geq m$ there exists a sequence $1 \leq k_1 < k_2 \dots \leq n$ such that $x_j = i_{k_j} \forall 1 \leq j \leq m$.

- **Definition 5** An sequence $t =< t_1, t_2, \ldots, t_m >$ is said to be "contained" in q-sequence $s =< s_1, s_2, \ldots, s_m >$ (i.e. $t \simeq s$). if $n \geq m$ there exists a sequence $1 \leq k_1 < k_2 \ldots \leq n$ such that $t_j \subseteq s_{k_j} \forall 1 \leq j \leq m$.

- **Definition 6** (utility of a sequence in s and D) Consider a sequence t and a q-sequence s, the utility of t in s, denoted as $u(t, s)$, may have different utility values. We choose the maximum utility among these utility values as the utility of t in s, as defined as $u(t, s) = max\{u(s_k)|t \simeq s_k \cap s_k \in s\}$. Let $u(t)$ denote the overall utility of a sequence t in a quantitative sequential database D. It is defined as $u(t) = \sum_{t \simeq s \cap s \in D} u(t, s)$.

- **Definition 7** (high-utility sequential pattern, HUSP) A sequence t is said to be a high-utility sequential pattern in D (denoted as HUSP) if its overall utility in D satisfies: $HUSP \leftarrow \{t|u(t) \geq \delta \cdot u(D)\}$, where δ is the minimum utility threshold δ (usually given as a percentage)

Problem statement: Given a quantitative sequential database D (with a profit-table), and a user-defined minimum utility threshold δ. The utility-driven mining problem of high-utility sequential pattern mining (HUSPM) consists of enumerating all HUSPs whose overall utility values in this database are no less than the given minimum utility account, such as $\delta \cdot u(D)$.

3 Projection-based Utility Mining on sequence data (ProUM)

3.1 Lexicographic Tree

The complete search space of HUSPM can be represented abstractly as the lexicographic sequence tree.

ProUM algorithm works by extending given sequence to some new sequence.

Given some sequence t, we can extend t by two ways:

- Add an item i_j to last itemset of t. denotes as $t + i_j$
- Add an item i_j to a new itemset appended after the last itemset of t, denoted as $t \oplus i_j$.

By this way start with empty sequence (root of tree) and recursively solve the problem for every node. But we still needs some pruning strategies at every stage of the recursion. Consider following Tree with two sequences, $< [a], [b], [a, b] >$ and $< [a], [a, b], [a] >$

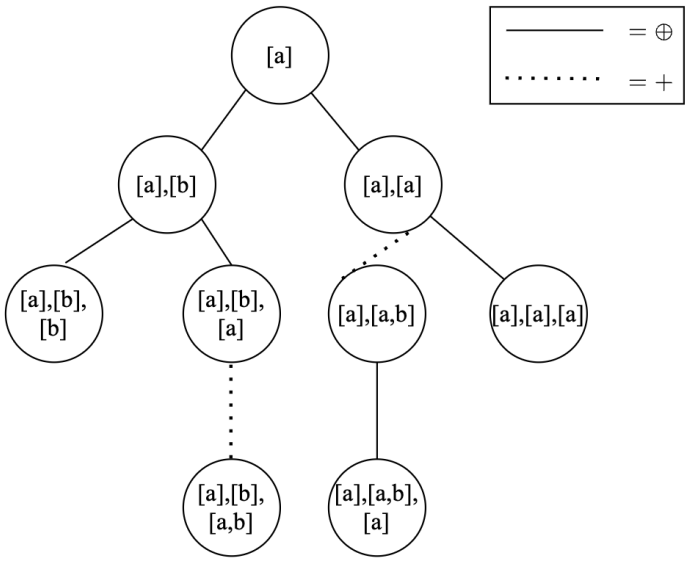

Figure 1: Lexicographic Tree with prefix [a]

3.2 Projection Mechanism

Remaining Utility: Given a sequence t and a sequence database D, the remaining utility of t in a q-sequence s is the overall utilities of all items which positions are after t in s, and defined as: $ru(t, s) = \max\{ru(t, p_k, s)\}$, where p_k is the k-position, and $ru(t, p_k, s) \sum_{i' \in s \cap t \leq i'} u(i')$. Thus, the overall remaining utility of t in D is defined as: $ru(t) = \sum_{s \in D} ru(t, s)$.

First matching: Given two q-sequences s and s,if s $\in$ s,the extension of s in s is said to be the rest of s after s, and is denoted as $< s - s >_{\text{rest}}$. Given a sequence t and a q-sequence s,if $t \simeq s_k \cap s_k \in s(t \in s)$, the rest of t in s is the rest part of s after s_k, which is denoted as $< s - t >_{\text{rest}}$, where s_k is the first match of t in s.

Let's say at any stage we have sequence t. Extending t by some item i_j. Let $t' = t + i_j$ or $t' = t \oplus i_j$.

- if t' is HUSP we can add it to set of HUSPs and continue it to next stage
- if t' is not HUSP, it doesn't guarantee that there does not exist any HUSP s which can be made from t by extension. But to reduce search space we can define a function to filter some sequences which will surely not gonna contribute to any HUSP in later stage. Hence if $ru(t, s) + u(t, s) \geq \delta \cdot u(D)$ we will continue with t, else just discard it.

This was the algorithm we analysed last semester, One can say this as "clever brute force". But during testing it was concluded that algorithm will fail to converge on some family of data sets.

4 Soft computing approach to the problem

4.1 Reduction to HUIM

This Problem looks very complex, and at very first moment there is nothing we can do except obvious brute force But we might ask is this problem similar to some problem we have seen in past. It turned out to be true, HUSPM is generalization of HUIM (High utility itemset mining) problem, in fact if we put constraint on length of sequence to be always 1, then it will reduced to HUIM. But in practise we need sequence with length greater than 1. Lets make some observations.

- Solve the HUSPM problem if length of sequence is always 1. We modified the ABC-HUIM algorithm to accomplish this.
- Now in any HUIM algorithm we get two types of Item sets as result. One which are High utility item-sets, Second one which currently are not high utility item-sets but they might be if we add or remove some items from it. For the problem of High utility sequential pattern mining we are interested in second type of item-sets. Because these are item-sets which we can "extend" to sequence of larger lengths.
- Now lets change the definition of word "extend" in HUSPM. In HUIM extension means adding/removing items from current set to make it HUI. In HUSPM extension of sequence would be adding/removing Item-sets to make it HUSP. At a time we wont change more than one item set. i.e. we will fix some sequence (Initially empty sequence) try to add/remove some itemset to make it HUSP.

4.2 Family of Soft Computing Algorithms for HUSPM

Using above observations we can design many types of algorithms based of internal HUIM model. Lets consider one such model.

- Define function f(t) on a sequence t as $f(t) = \sum_{t \simeq s \cap s \in D} u(s)$. Define IsPromising(s) on a sequence s as

$$IsPromising(s) = \begin{cases} 1 & \text{if } f(s) \geq \delta * u(D) \\ 0 & \text{else} \end{cases}$$

- Let PromisingSeq be the set of promising sequences. Initially would only contains empty sequence.
- At each iteration of algorithm randomly pick some sequence t from PromisingSeq (We can use roulette wheel method). Now Our goal is to improve u(t) and f(t) by adding one itemset at the end of t. Here we can use HUIM algorithm. Using HUIM algorithm we can get some candidate item-sets which can improve u(t) and f(t). If no candidate item-set is found for t after some fixed amount of steps (Max Trials), we will discard t, else add $t \oplus X$ to the Promising Seq. (where X is some candidate item-set)

Above algorithm looks very promising and far less complicated than previous one. But There is still more work to be done on this strategy, like we can still improve f(t) by make upper bound more strict.

5 Conclusion and Future Work

5.1 Future Work

High Utility Sequential Pattern Mining (HUSPM) is an emerging field with rapidly growing applications and opportunities. Future expectations and work should include:

1. Implementing the second part of the algorithm that is using 1-HUSP's to generate the variable length HUSP's.
2. Further improvements in the soft computing approach by pruning and/or by using different data structures like utility linked lists etc. The approach is flexible, so it can be modified by slight twists in logic or by incorporating variants of the existing algorithms to produce better results.
3. Negative utility, Same algorithms can also be applied to negative utility but we have to change the upper bound strategies as current closure assume utility values are positive.

5.2 Knowledge Acquired

1. Learnt about HUSPM and implemented project based utility mining algorithm on sequence data
2. Learnt data pruning and other optimization techniques to improve performance
3. Soft computing to efficiently solve the problem of HUSPM

5.3 Conclusion

ProUM is definitely better than previous proposed algorithms but like previous algorithms ProUM would fail on Datasets with large transaction size. But as we proved above HUSPM is generalized version of HUIM, and so much works has already been done in HUIM. We can reduce problem to HUSPM to Multiple sub problems of HUIM. Theoretically given soft computing approach should perform better than previous proposed algorithms for HUSPM.

References

[1] Wensheng Gan, Jerry Chun-Wei Lin, Jiexiong Zhang, Han-Chieh Chao, Hamido Fujita, and Philip S. Yu. Proum: Projection-based utility mining on sequence data. *Information Sciences*, 513:222–240, 2020.

[2] Rakesh Agrawal, Ramakrishnan Srikant, et al. Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, volume 1215, pages 487–499. Citeseer, 1994.

[3] Sebastián Ventura, José María Luna, et al. *Pattern mining with evolutionary algorithms*. Springer, 2016.

[4] Rakesh Agrawal and Ramakrishnan Srikant. Mining sequential patterns. In *Proceedings of the eleventh international conference on data engineering*, pages 3–14. IEEE, 1995.

[5] Chunkai Zhang, Zilin Du, Wensheng Gan, and Philip S. Yu. Tkus: Mining top-k high utility sequential patterns. *Information Sciences*, 570:342–359, 2021.