High Utility Itemset Mining from Transactional Databases

Report submitted in fulfillment of the requirements
for the Exploratory Project of

Second Year IDD

by

Vaibhav Khater

Under the guidance of

Dr. Bhaskar Biswas

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

Dedicated to

Dr. Bhaskar Biswas, Associate Professor, Department of Computer Science and Engineering, IIT (BHU) Varanasi, for invaluable guidance and support for completion of this project .

Declaration

I certify that

1. The work contained in this report is original and has been done by myself and the general supervision of my supervisor.

2. The work has not been submitted for any project.

3. Whenever I have used materials (data, theoretical analysis, results) from other sources, I have given due credit to them by citing them in the text of the thesis and giving their details in the references.

4. Whenever I have quoted written materials from other sources, I have put them under quotation marks and given due credit to the sources by citing them and giving required details in the references.

Place: IIT (BHU) Varanasi
Date: 20/4/2023

Vaibhav Khater
B.Tech. or IDD Student
Department of Computer Science and Engineering,
Indian Institute of Technology (BHU) Varanasi,
Varanasi, INDIA 221005.

Certificate

*This is to certify that the work contained in this report entitled "**High Utility Itemset Mining from Transactional Databases**" being submitted by **Vaibhav Khater** (**Roll No. 21074033**), carried out in the Department of Computer Science and Engineering, Indian Institute of Technology (BHU) Varanasi, is a bona fide work of our supervision.*

Place: IIT (BHU) Varanasi
Date: 20/4/2023

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

Acknowledgments

I would like to express my sincere gratitude to Dr. Bhaskar Biswas, Associate Professor, Department of Computer Science and Engineering, IIT (BHU) Varanasi for hir valuable insights into this project.

Place: IIT (BHU) Varanasi

Date: 20/4/2023 **Vaibhav Khater**

Abstract

The discovery of high utility itemsets (HUI's) from transactional databases has found its immense application in various industries like Market Basket Analysis ,Web Click Stream mining etc. In this project I have implemented algorithms to find Transactional Weighted High Utility Itemsets (TWHUI's), mining Short Period High Utility Itemsets (SPHUI's) designed to identify patterns in a transactional database that appear regularly, are profitable, and also yield a high utility under the period constraint both with effective pruning strategies. The aim of discovering short-period high-utility itemsets (SPHUI) is hence to identify patterns that are interesting both in terms of period and utility. The paper implements a baseline two-phase short-period high-utility itemset (SPHUITP) mining algorithm to mine SPHUIs in a level-wise manner. Then, to reduce the search space of the SPHUITP algorithm and speed up the algorithm, two pruning strategies are developed and integrated in the baseline algorithm. The resulting algorithms are denoted as (SPHUI-MT) and (SPHUI-TID).

Contents

Chapter 1

Introduction

1.1 Overview

Data mining techniques are employed to extract information from exceedingly huge databases and transform it into a straightforward, comprehensive structure for future use. It has various applications inn fields like Market basket analysis [1,5],Web Click stream mining [1,6] and Web Mining [2,7]. The key restrictions of traditional algorithms include the fact that each item is given equal weight and cannot appear more than once in a single transaction. But these presumptions frequently do not hold true in real-world circumstances. To overcome this , the utility is defined as the measurement of the usefulness of an itemset. Utility values of items in a transaction database consist of two parts: the item profit (external utility) and the quantity of the item in one transaction (internal utility). The utility of an itemset is defined as the external utility multiplied by the internal utility However, it is very challenging to mine High Utility Itemsets because they do not hold the Downward Closure Property (DCP),i.e.a superset of low utility itemsets may be high utility itemsets in the HUIM approach. This work is based on implementation of mining Transactional Weighted High Utility Itemsets (TWHUI's)and mining Short Period High Utility Itemsets (SPHUI's) designed to identify patterns in a transactional database

that appear regularly, are profitable, and also yield a high utility under the period constraint both with effective pruning strategies. The aim of discovering short-period high-utility itemsets (SPHUI) is hence to identify patterns that are interesting both in terms of period and utility. The paper implements a baseline two-phase short-period high-utility itemset (SPHUITP) mining algorithm to mine SPHUIs in a level-wise manner. Then, to reduce the search space of the SPHUITP algorithm and speed up the algorithm, two pruning strategies are developed and integrated in the baseline algorithm. The resulting algorithms are denoted as (SPHUI-MT) and (SPHUI-TID).

1.2 Motivation of the Research Work

I have been interested in field of learning and implementing new algorithms as well as their applications and analyzing their behaviour. This field of data mining particular interested me because of its increasing application in modern era and the valuable assets it adds to the industry. I have also been working and skilled in efficient bitmasking, pruning, memorisation strategies and competitive programming, and believed my skills could find the usecase here for better and fast implementations. It is also very astonishing to feel that how such large databases can be curated to find the higher utility subsets with the help of mining algorithms.

1.3 Organisation of the Report

The work focuses on presenting :

1.) Extracting Transactional Weighted High Utility Itemsets (TWHUI's) from transactional databases.

2.) To find short-period high-utility itemsets (SPHUIs) in a level-wise manner, an effective two-phase short-period high-utility itemset mining (SPHUITP) approach is presented. The technique uses an upper-bound on the utility of itemsets in its initial

phase to identify a limited number of potential SPHUIs. The algorithm then determines the actual SPHUIs in this group of candidates in the second phase.

3) To speed up the baseline SPHUITP algorithm, two pruning strategies are further introduced. The resulting algorithms are respectively named SPHUIMT and SPHUITID. The two strategies are designed to reduce the search space for mining SPHUIs.

Chapter 2

Project Work

Transaction Weighted High Utility Itemset Mining (TWHUI's):

This section describes the startegy used for mining Transactional Weighted High Utility Itemsets from a given database. Before further proceeding, you are required to know basic terminologies and definitions related to transactional databases[3].

Problem statement: Given a transaction database D and a user-specified minimum utility threshold δ, the objective of the HUIM is to discover the complete set of candidates that have utilities no less than the minimum utility count ($\delta \times$ tu(D)).

It is a challenging task to prune the search space in the HUIM because the downward closure property does not hold for the utility measure of an item in the database. To address this issue, the TWDCP(Transaction Weighted Downward Closure Property) [3] is used along with TWU Pruning Strategies as describes further.
The transaction database consists eight transactions t1, t2, . . . , t8 as shown in Table 1 that will be used as a running example. It consists five items a, b, c, d, e. The external utility of each item is shown in Table 2. For example, the transaction t3 consists four items a, b, d, e and has internal utility 3, 1, 5, 2 respectively. The

external utility of a, b, c, d, e are 3, 2, 1, 4, 1 respectively.

tid	transaction
t_1	(b,1), (c,1), (d,3)
t_2	(a, 2), (c, 4), (e, 1)
t_3	(a, 3), (b, 1), (d, 5), (e, 2)
t_4	(b, 2), (c, 1), (e, 3)
t_5	(a, 2)
t_6	(b, 1), (d, 2)
t_7	(a, 3), (b, 2), (d, 4), (e, 1)
t_8	(a, 3), (d, 1), (e, 3)

Table 2.1 Transaction Database.

Items	a	b	c	d	e
External Utility	3	2	1	4	1

Table 2.2 External Utility Value.

TWU Based Pruning Strategy :

1.Overestimation : The TWU of an itemset X is higher than or equal to its utility.

2.Antimonotone : If X itemset is a subset of Y itemset then TWU(X) is larger than TWU(Y).

3.Pruning : If Transaction Weighted Utility TWU(x) is larger or equal to minimum utility count, it can be considered as a high utility itemset , otherwise not.

Short Period High Utility Itemsets Mining (SPHUIM's):

This section describes the startegy used for mining Short Period High Utility Itemsets from a given database. Before further proceeding, you are required to know basic terminologies and definitions related to transactional databases[ ].

Problem statement: Given a transactional database D, let there be a user-provided

minimum utility threshold δ and a maximal period threshold θ. The goal of short-period high-utility itemset mining (SPHUIM) is to discover the set of itemsets in which each pattern satisfies the conditions: discover the complete set of candidates that have utilities no less than the minimum utility count ($\delta \times$ tu(D)) and maximum period is lesser than ($\theta \times$ n) where n is the number of transactions

The suggested approach is intended to find highly useful patterns with frequent occurrences over brief time periods. The suggested framework takes into account both the short-period and utility restrictions to identify itemsets that are highly useful and frequently present in databases. To find SPHUIs in a level-wise manner, an effective two-phase short-period high-utility itemset mining (SPHUITP) approach is described. The technique uses an upper-bound on the utility of itemsets in its initial phase to identify a limited number of potential SPHUIs. The algorithm then determines the actual SPHUIs in this group of candidates in the second phase. Two pruning strategies are also introduced to speed up the baseline SPHUITP algorithm. The resulting algorithms are called SPHUIMT and SPHUITID. Two strategies are designed to reduce the search space for mining SPHUI.

TID	A	B	C	D	E
1	5	15	0	2	0
2	0	0	0	1	1
3	0	1	0	0	1
4	6	0	0	1	1
5	0	1	0	1	0
6	1	0	0	1	0
7	5	1	0	1	0
8	0	0	12	0	2
9	8	0	0	3	0
10	0	10	10	1	1

Table 2.3 Transactional Database.

To mine, in the first phase of two phase algorithm all HTWSPUI(High Transac-

Items	A	B	C	D	E
Profit	5	4	6	2	10

Table 2.4 Profit table.

tion Weighted Utility Short Period Itemsets) are mined successfully from the given data set as it can be proved that all SPHUI itemsets are subset of HTWSPUI[4]. The mining is based on TWSPDC(Transaction Weighted Short Period Downward Closure Property) which states if an itemset X is a HTWSPUI its all subsets are also HTWSPUI and if X is not a HTWSPUI then all its supersets are not SPHUI.

In the second phase of the implementation we can individually check all the valid candidates from the first phase and check if it satisfies both the SPHUI constraints.This can be improved further by using two pruning strategies SPHUIMT and SPHUITID.

SPHUIMT : If MT(X) is larger than the maximal period value , then itemset can be successfully discarded from a valid candidate for SPHUI.

SPHUITD : If the utility of a transaction is less than the minimum utility value , the transaction ID (Tid) of the respective transition can be removed from the Tid-lists of all itemsets.

Chapter 3

Results and Discussions

Transaction Weighted High Utility Itemset Mining (TWHUI's):

Mining TWHUI's from dataset given in table 2.1, 2.2 gives the following results :

Itemsets (TWHUI)	Utility
a	96
b	96
a,b	63
c	34
b,c	23
d	104
a,d	79
b,d	88
a,b,d	63
c,d	15
b,c,d	15
e	98
a,e	90
b,e	71
a,b,e	63
c,e	19
d,e	79
a,d,e	79
b,d,e	63
a,b,d,e	63

Table 3.1 Mined TWHUI's from Table 2.1, 2.2

Table 3.1 shows the successfully mined High Transaction Weighted Utility Item-

sets from dataset present in Table 2.1, 2.2 for which minimum utility count came out to be 12.9 for threshold value δ being set to 0.1.

Short Period High Utility Itemsets Mining (SPHUIM's):

We analyze the following results for dataset given in Table 2.3, 2.4 for δ=0.01 and θ=0.3.

(i)SPHUI-TP:

Successfull candidates to make through first phase (HTWSPUI's):

Itemsets (HTWSPUI)	Utility
A	215
B	252
C	204
D	345
E	272
A,B	120
A,D	215
B,C	112
B,D	238
B,E	126
C,D	112
C,E	204
D,E	166
A,B,D	120
B,C,D	112
B,C,E	112
B,D,E	112
C,D,E	112
B,C,D,E	112

Table 3.2 Mined HTWSPUI's from Table 2.3, 2.4

SPHUI's after second phase :

Out of the 19 possible candidates given in Table 3.2 , we mine the actual Short Period

High Utility Itemsets(SPHUI's):

Itemsets (SPHUI)	Utility
A	215
A,D	215

Table 3.3 Mined SPHUI's from Table 2.3, 2.4

(ii)SPHUI-MT:

Now we analyze on the same dataset with MT-pruning strategy combined.

Successfull candidates to make through first phase (HTWSPUI-MT's):

Itemsets (HTWSPUI-MT)	Utility
A	215
B	252
D	345
A,D	215

Table 3.4 Mined HTWSPUI's from Table 2.3, 2.4

SPHUI's after second phase :

Out of the 4 possible candidates given in Table 3.2 , we mine the actual Short Period High Utility Itemsets(SPHUI's):

Itemsets (SPHUI)	Utility
A	215
A,D	215

Table 3.5 Mined SPHUI's from Table 2.3, 2.4

(iii)SPHUI-TID:

Now we analyze on the same dataset with TID-pruning strategy combined.

Successfull candidates to make through first phase (HTWSPUI's):

Itemsets (HTWSPUI-TID)	Utility
A	215
D	345
A,D	215

Table 3.6 Mined HTWSPUI-TID's from Table 2.3, 2.4

SPHUI's after second phase :

Out of the 3 possible candidates given in Table 3.2 , we mine the actual Short Period High Utility Itemsets(SPHUI's):

Itemsets (SPHUI)	Utility
A	215
A,D	215

Table 3.7 Mined SPHUI's from Table 2.3, 2.4

Analysis and Comparision of all three strategies :

We can see that SPHUI mined from all the three strategies are consistent with each other.But we can note that the number of candidates to check for SPHUI in the final stage are different for all three algorithms. They are $19, 4 and 3$ for SPHUI-TP, SPHUI-MT and SPHUI-TID respectively which also imply their efficiency on larger datasets.

Runtime Analysis :

We plot a graph on runtime vs minimum utility threshhold for SPHUI, SPHUIMT and SPHUITID on a dataset randomly generated having 15 different items and 100 transactions with MP fixed as 0.3.

It can be seen that both the SPHUIMT and SPHUITID algorithms outperform the baseline SPHUITP algorithm. It can be seen that order of growth of SPHUITP algorithm is significantly greater than both SPHUIMT algorithm and SPHUITID al-

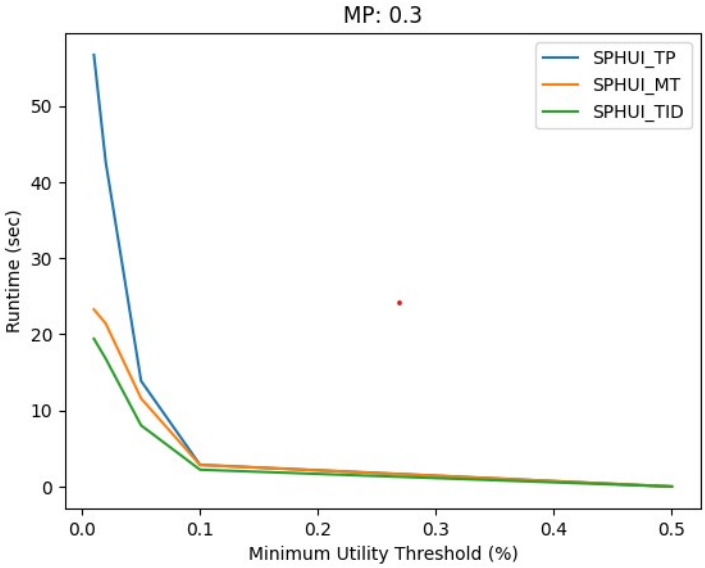

Figure 3.1 Runtime Analysis (fixed θ)

gorithm.

Now, we plot a graph on runtime vs maximal period threshhold for SPHUI, SPHUIMT and SPHUITID on a dataset randomly generated having 15 different items and 100 transactions with MU fixed as 0.002.

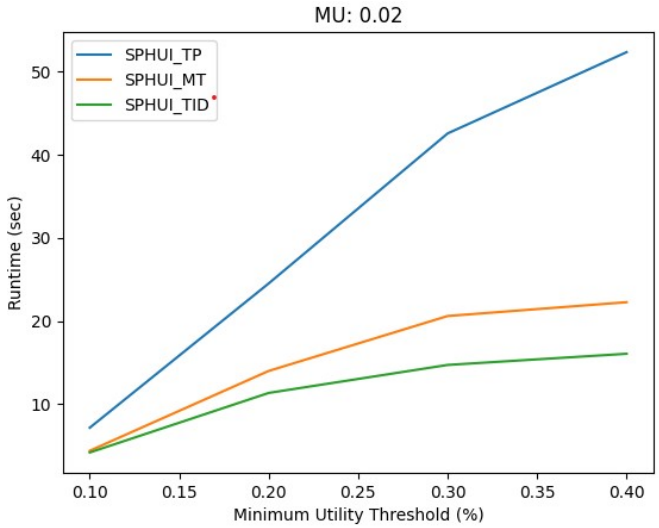

Figure 3.2 Runtime Analysis (fixed δ)

It can again be seen that both the SPHUIMT and SPHUITID algorithms outper-

form the baseline SPHUITP algorithm. It can again be seen that order of growth of SPHUITP algorithm is significantly greater than both SPHUIMT algorithm and SPHUITID algorithm.

Chapter 4

Bibliography

[1] Chowdhury Farhan Ahmed, Syed Khairuzzaman Tanbeer, Byeong-Soo Jeong, and Young-Koo Lee. Efficient tree structures for high utility pattern mining in incremental databases. IEEE Transactions on Knowledge and Data Engineering, 21(12):1708–1721,2009.

[2] Chowdhury Farhan Ahmed, Syed Khairuzzaman Tanbeer, and Byeong-Soo Jeong. A framework for mining high utility web access sequences. IETE Technical Review, 28(1):3–16, 2011.

[3] Kumar, Rajiv; Botho University, Faculty of Engineering and Technology SINGH, KULDEEP; University of Delhi, Department of Computer Science,High Utility Itemsets Mining from Transactional Databases: A Survey, The Knowledge Engineering Review:7-17,2022.

[4] Jerry Chun-Wei Lin, Jiexiong Zhang , Philippe Fournier-Viger, Tzung-Pei Hong , Ji Zhang ,A two-phase approach to mine short-period high-utility itemsets in transactional databases : Advanced Engineering Informatics:8-2017.

[5] Wensheng Gan, Jerry Chun-Wei Lin, Philippe Fournier-Viger, Han-Chieh Chao, and Hamido Fujita. Extracting non-redundant correlated purchase behaviors by utility measure. Knowledge-Based Systems, 143:30–41, 2018.

[6] Hua-Fu Li, Hsin-Yun Huang, Yi-Cheng Chen, Yu-Jiun Liu, and Suh-Yin Lee. Fast and memory efficient mining of high utility itemsets in data streams. In 2008 eighth IEEE international conference on data mining, pages 881–886. IEEE, 2008

[7] Brijesh Bakariya and G.S. Thakur. An efficient algorithm for extracting high utility itemsets from weblog data. IETE Technical Review, 32(2):151–160, 2015.