Mining Top-k High Utility Patterns Over Data Streams

Morteza Zihayat and Aijun An

Technical Report CSE-2013-09

March 21 2013

Department of Computer Science and Engineering
4700 Keele Street, Toronto, Ontario M3J 1P3 Canada

Mining Top-k High Utility Patterns Over Data Streams

Morteza Zihayat[a], Aijun An[a]

{*zihayatm,aan*}*@cse.yorku.ca*

[a]*Department of Computer Science & Engineering,York University, Toronto, Canada*

Abstract

Online high utility itemset mining over data streams has been studied recently. However, the existing methods are not designed for producing top-k patterns. Since there could be a large number of high utility patterns, finding only top-k patterns is more attractive than producing all the patterns whose utility is above a threshold. A challenge with finding top-k high utility itemsets over data streams is that it is not easy for users to determine a proper minimum utility threshold in order for the method to work efficiently. In this paper, we propose a new method for finding top-k high utility patterns over sliding windows of a data stream. The method (named T-HUDS) is based on a compressed tree structure, called *HUDS-tree*, that can be used to efficiently find potential top-k high utility itemsets over sliding windows. *T-HUDS* uses a new utility estimation model to more effectively prune the search space. We also propose several strategies for initializing and dynamically adjusting the minimum utility threshold. We prove that no top-k high utility itemset is missed by the proposed method. Our experimental results on real and synthetic datasets show that our strategies and new utility estimation model work very effectively and that *T-HUDS* outperforms two state-of-the-art high utility itemset algorithms substantially in terms of execution time and memory storage.

Keywords:
High utility pattern mining, data stream, top-k pattern mining, sliding window

1. Introduction

Frequent pattern mining is an important task in data mining and has been extensively studied by many researchers [1, 12, 11]. Given a data set of transactions, each containing a set of items, frequent pattern mining is to find a set of itemsets whose support (i.e., the number of transactions containing the itemset) is no less than a minimum support count. However, in frequent pattern mining, the number of occurrences of an item inside a transaction is ignored in the problem setting, so is the importance (such as price or weight) of an item in the data set. In practice, some items or itemsets with low support in the data set may bring high profits due to their high prices or high frequencies inside transactions. Obviously, identifying such itemsets with high profits is important for business planning and operation. However, such itemsets may be missed by frequent pattern mining.

In view of this, high utility itemset mining has been studied recently [3, 4, 25, 18]. An itemset is a *high utility itemset (HUI)* if its utility (such as the total profit that the itemset brings) in a data set is no less than a minimum utility threshold. Finding high utility itemsets has been considered to be important in various applications, such as retail marketing, web click analysis, and biological gene analysis [16, 15, 2]. However, mining HUIs is not as easy as mining frequent itemsets. This is due to the fact that the utility of an itemset does not have the downward closure property, which would allow effective pruning of search space during the HUI mining process. To deal with such a challenge, most of the HUI mining methods use an overestimated utility, called *transaction weighted utility (TWU)* (to be defined in section 2), to first find itemsets whose TWU is no less than the minimum utility threshold (called high TWU itemsets) and then compute the exact utilities of high TWU itemsets to identify those whose utility satisfies the minimum utility threshold. The benefit of using TWU is that TWU has the downward closure property, which allows the use of Apriori-like or FP-growth-like algorithms in the first phase of HUI mining to efficiently find high TWU itemsets.

Since data streams have become widespread in many fields, such as sensor network monitoring, trade management, and medical data analysis, methods for mining HUIs from data streams have been proposed [2, 16, 15, 24]. In comparison to static data, data streams have some unique properties, such as very fast data arrival rate, unknown or unbounded size of data and inability to backtrack over previously arriving transactions. To deal with such

challenges, a HUI mining method in [2] (named *HUPMS*) uses a compact data structure similar to FP-tree [12] to compress the transactions in the data set and uses a pattern growth method (similar to FP-growth) to efficiently identify all the high TWU itemsets with respect to a minimum utility threshold. HUIs are then identified from the set of high TWU itemsets after scanning the recent data in a sliding window for the second time to compute the exact utility of these itemsets. Although the use of TWU allows effective pruning of the search space due to its downward closure property, it is a very loose estimate of the true utility of an itemset. As a result, the number of high TWU itemsets found in the first phase of the method can be high and many of them do not satisfy the minimum utility threshold. Thus, the overall time for finding HUIs can be too long to satisfy the fast data processing requirement for data streams.

Another problem with the method in [2] and many other HUI mining methods is that the user needs to supply a minimum utility threshold. However, it is often difficult for the user to specify a minimum utility threshold, especially if the user has no background knowledge in the application domain. If the threshold is set too low, a large number of HUIs can be found, which is not only time and space consuming but also makes it hard to analyze the mining results. On the other hand, if the threshold is set too high, there may be very few or even no HUIs being found, which means that some interesting patterns are missed. Figure 1 illustrates the major impact of setting the threshold on the run time and the number of obtained HUIs on *BMS-POS* data set [9] by the state-of-the-art *HUPMS* algorithm [2][1]. For example, decreasing the threshold from 1.1% into 1.0%, increases the run time and the number of candidates significantly (i.e., by almost 5 times in the run time and by 36 times in the number of candidates).

A solution to this threshold setting problem is to mine top-k high utility itemsets, in which the user supplies k, the number of HUIs to be returned. A benefit of mining top-k patterns is that it is easier and more intuitive for the user to indicate how many patterns they would like to see than specifying a utility threshold. In addition, the number of returned patterns will be under control and the result will not overwhelm the user. A method for top-k HUI

[1] A similar example on a different data set was used in [27] to illustrate the threshold setting impact. However, in our example, the problem is more obvious, and the data set is treated as a data stream. The values in the graphs are the average over all the sliding windows.

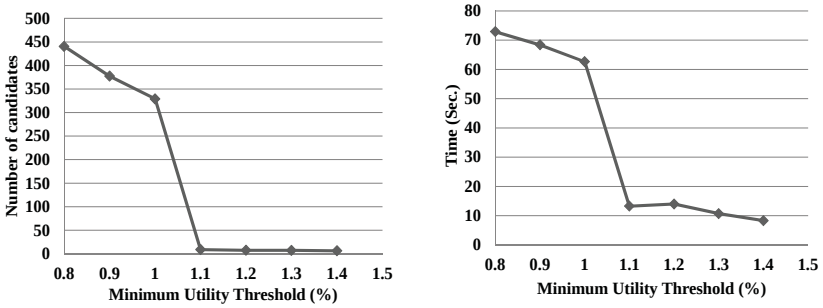


Figure 1: Run time and number of obtained high utility itemsets in BMS-POS data set for different minimum utility threshold values (Window size = 3 Batches)

mining was proposed very recently in [27]. The method is designed for static data, not for data stream mining. A major challenge in top-k HUI mining is that the number of itemsets is exponential and it is infeasible to compute the utilities of all the itemsets and identify the top-k ones. A minimum utility threshold is thus needed in the mining process to prune the search space. The method proposed in [27] initializes the threshold to 0 or the kth highest value of the lower bounds for the utility of certain 2-itemsets, and then gradually raise the threshold during the mining process to prune the search space. The authors proposed a few strategies for raising the threshold. However, their initial threshold is too low and can lead to generation of a large number of potential HUIs in the first phase of the method. In addition, their method is not designed for data streams.

In this paper, we propose more effective strategies for automatically initializing and dynamically adjusting the minimum utility threshold for mining top-k high utility itemsets over data streams. Three of our strategies can be applied to both static and streaming data, and one of them is specially designed for data streams. We use a sliding window based data stream mining method, in which a set of recent data (called a *sliding window*) is the target of mining. A sliding window consists of a fixed number of most recent batches, each batch containing a set of transactions. When a new batch arrives, the sliding window moves forward to include the new batch and at the same time remove the oldest batch if the maximum number of batches in the window has been reached before the new batch comes. In addition to the new strategies for setting and adjusting the threshold, we also propose to use another over-estimated utility as the search heuristic for finding HUIs in the

first phase of the top-k HUI mining process. This over-estimate (called *prefix utility*) is more effective than the most commonly used TWU in pruning the search space because it is a closer estimate of the true utility than TWU. The contributions of the paper are as follows:

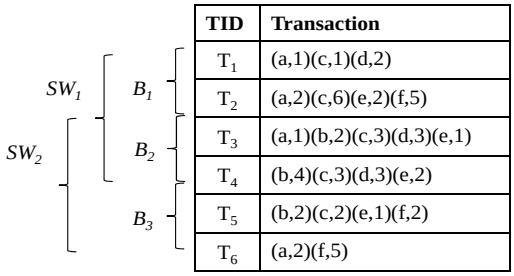
- We are the first to propose a method for mining top-k high utility itemsets from data streams. To the best of our knowledge, existing methods for mining HUIs over data streams do not address the issue of mining top-k HUIs, and previous top-k HUI mining methods do not work on data streams.
- We propose several strategies for initializing and dynamically adjusting the minimum utility threshold during the top-k HUI mining process. We prove that using these strategies will not miss any top-k HUIs.
- We propose an over-estimate of the itemset utility, which is closer to the true utility than TWU. We prove that this estimate (i.e., *prefix utility*) has a special type of downward closer property, which allows it to be used in the pattern growth method to effectively prune the search space. Using a closer over-estimate results in fewer candidates being generated in the first phase of the method.
- We propose an FP-tree-like compact data structure (called *HUDS-tree*) to store the information about the transactions in a sliding window. The tree is used to compute the prefix utility and to initialize and adjust the minimum utility threshold.
- We conduct an extensive experimental evaluation of the proposed method on both real and synthetic data sets, which shows that our proposed method is faster and less memory consuming than the state-of-the-art methods.

The paper is organized as follows. Preliminary definitions and a problem statement are given in Section 2. In Section 3, we describe the challenges in solving our problem and define some concepts used in our methods. In Section 4, we present the *HUDS-tree* structure and our algorithms for finding top-k *HUIs*. The experimental results are presented in Section 5. Related work is discussed in Section 6. In Section 7 we conclude the paper.

2. Preliminaries and Problem Statement

Let $I = \{i_1, i_2, ..., i_m\}$ be a set of items and each item $i_j \in I$ is associated with a positive number $p(i_j)$, called its *external utility* (which can be the price or profit) of item i_j. Let D be a set of N transactions: $D = \{T_1, T_2, ..., T_N\}$

Figure 2: Example of transaction data base and external utility of items



TID	Transaction
T_1	(a,1)(c,1)(d,2)
T_2	(a,2)(c,6)(e,2)(f,5)
T_3	(a,1)(b,2)(c,3)(d,3)(e,1)
T_4	(b,4)(c,3)(d,3)(e,2)
T_5	(b,2)(c,2)(e,1)(f,2)
T_6	(a,2)(f,5)

Item Name	a	b	c	d	e	f
External utility	3	6	5	8	4	3

such that for $\forall T_j \in D, T_j = \{(i, q(i, T_j))|i \in I, q(i, T_j)$ is the quantity of item i in transaction $T_j\}$. Figure 2 shows an example of a data set with six transactions.

Definition 1. ***Utility of an item*** *i* ***in a transaction*** *T_j is defined as: $u(i, T_j) = q(i, T_j) \times p(i)$ where $q(i, T_j)$ is the* quantity *of item i in transaction T_j and $p(i)$ is* external utility *of item i.*

Definition 2. ***Utility of an itemset*** *X* ***in a transaction*** *T_j is defined by: $u(X, T_j) = \sum_{i \in X} u(i, T_j)$.*

For example, $u(\{bc\}, T_3) = 2 \times 6 + 3 \times 5 = 27$ in Figure 2.

Definition 3. ***Utility of an itemset*** *X* ***in a data set*** *D of transactions is defined as: $u_D(X) = \sum_{X \subseteq T_j \wedge T_j \in D} \sum_{i \in X} u(i, T_j)$.*

We use $u(X)$ to denote $u_D(X)$ when data set D is clear in the context.

Definition 4. ***Utility of a transaction*** *T_j is denoted as $TU(T_j)$ and computed as $u(T_j, T_j)$.*

Definition 5. ***(High Utility Itemset (HUI))*** An itemset X is called a high utility itemset (HUI) *on a data set D if and only if $u_D(X) \geq min\_util$ where* min_util *is called a* minimum utility threshold.

A challenge in mining high utility itemsets (HUIs) is that the utility of an itemset does not have the downward closure (i.e., anti-monotone) property. That is, the utility of an itemset does not decrease monotonically when adding items to the itemset. It changes irregularly. Thus, unlike in frequent itemset mining, we cannot use the utility of an itemset to prune the search space in high utility itemset mining because a superset of a low utility itemset may be a high utility itemset.

To solve this problem, an over-estimated utility of an itemset (instead of the exact utility) is commonly used in the HUI mining process to prune the search space. Most of the recent methods use *transaction-weighted utility (TWU)* as the over-estimated utility:

Definition 6. ***Transaction-Weighted Utility (TWU)*** *of an itemset X over a data set D is defined as:* $TWU_D(X) = \sum_{X \subseteq T_j \bigwedge T_j \in D} TU(T_j)$.

Clearly, $TWU_D(X) \geq u_D(X)$. In addition, TWU satisfies the downward closure property, that is, for all $Y \subseteq X$, $TWU_D(Y) \geq TWU_D(X)$. Thus, most of the HUI mining methods (e.g., [16, 2]) uses the TWU values of the itemsets to prune the search space. That is, they find all the itemsets whose TWU is no less than the minimum utility threshold. Since $TWU_D(X)$ is an overestimate of $u_D(X)$, the procedure does not miss any high utility itemset. But the true utility of a generated itemset may be lower than the minimum utility threshold. Thus, these methods use a second phase to compute the exact utility of the generated itemsets and remove those whose utility is lower than the threshold.

We are interested in mining top-k HUIs in data streams. In a data stream environment, transactions come continually over time, and they are usually processed in batches. A **batch** B_i consists of transactions arriving continuously in a time period, i.e., $B_i = \{T_j, T_{j+1}, ..., T_m\}$. For example, assuming that the dataset in Figure 2 is a data stream and that each batch contains 2 transactions, there are three batches in the stream: $B_1 = \{T_1, T_2\}, B_2 = \{T_3, T_4\}$, and $B_3 = \{T_5, T_6\}$.

A **sliding window** consists of m most recent batches, where m is called the size of window, denoted as *winSize*. If the first batch in a sliding window is B_i, the window can be represented as $SW_i = \{B_i, B_{i+1}, ..., B_{i+winSize-1}\}$. As a new batch forms up in a data stream, the sliding window removes its oldest batch and adds the new batch to the window. For example, consider the data stream in Figure 2. Assume that the $winSize$ is 2. The first two batches form

the first sliding window: $SW_1 = \{B_1, B_2\}$. When the third batch B_3 is filled up with transactions, the second sliding window is formed: $SW_2 = \{B_2, B_3\}$. Data stream mining over sliding windows is to mine patterns from each new window once a new batch is added into the new window and the oldest batch is removed from the window. The problem tackled in the paper is defined as follows.

Problem 1. For each sliding window SW_i in a data stream, the problem is to find the top-k high utility itemsets in SW_i, ranked in descending order of their utility, where k is a positive integer given by the user.

3. Challenges and New Definitions

There are inherent challenges in mining top-k *HUIs* in data streams. First, since streaming data can come continuously in a high speed, they need to be processed as fast as possible. As mentioned earlier, the utility of an itemset does not have the downward closure property, and thus most of the existing HUI mining methods use TWU (an over-estimate of the itemset utility) as the search heuristic to prune itemsets whose TWU is below the minimum utility threshold. To further speed up the HUI mining process, we define another over-estimated utility of an itemset, which provides a closer estimation of the true utility of an itemset than TWU. This over-estimated utility, called *Prefix Utility*, is used in our HUI mining to more effectively prune the search space.

Definition 7. ***Prefix Utility of an itemset*** X ***in a transaction*** T. *Assume the items in* T *are ranked in an order (such as the lexicographic order) and that* $X \subseteq T$. *The* prefix set *of* X *in* T, *denoted as* $PrefixSet(X, T)$, *consists of all the items in* T *that are not ranked after any item in* X. *The* prefix utility *of* X *in* T *is defined as:*

$$PrefixUtil(X, T) = \sum_{i \in PrefixSet(X,T)} u(i, T)$$

Example 1. *In Figure 2, the prefix set of itemset* $\{ac\}$ *in transaction* T_3 *is* $\{abc\}$. *Thus,*

$$\begin{aligned} PrefixUtil(\{ac\}, T_3) &= u(a, T_3) + u(b, T_3) + u(c, T_3) \\ &= 3 + 12 + 15 = 30 \end{aligned}$$

Definition 8. ***Prefix Utility of an itemset*** X ***in a dataset*** D *is defined as:*

$$PrefixUtil_D(X) = \sum_{X \subseteq T_j \bigwedge T_j \in D} PrefixUtil(X, T_j)$$

Here we assume that items in all the transactions are ranked in the same order.

Example 2. *Let* D *be the dataset in Figure 2. Since only* T_1*,* T_2 *and* T_3 *in* D *contain itemset* $\{ac\}$*, we have*

$$\begin{aligned} PrefixUtil_D(\{ac\}) &= PrefixUtil(\{ac\}, T_1) + PrefixUtil(\{ac\}, T_2) + \\ &\quad PrefixUtil(\{ac\}, T_3) \\ &= 8 + 36 + 30 = 74 \end{aligned}$$

Property 1. *For any itemset* X *in a dataset* D*, the following holds:*

$$TWU_D(X) \geq PrefixUtil_D(X) \geq u_D(X)$$

Lemma 1. *Assume that items in all the transactions in a dataset* D *are ranked in an order. Let* X *be an itemset and* $X = Y \cup \{i\}$ *where* i *is the last item in* X *in the ranked order. For all* $Z \subseteq Y$*,*

$$PrefixUtil_D(Z \cup \{i\}) \geq PrefixUtil_D(X)$$

.

Proof 1. *Let* S_X *be the set of transactions containing* X *in a data set* D*. According to Definition 8, we have*

$PrefixUtil_D(Z \cup \{i\}) = PrefixUtil_{S_X}(Z \cup \{i\}) + PrefixUtil_{D-S_X}(Z \cup \{i\})$.

Since itemset $Z \cup \{i\}$ *contains the last item in* X *and* $Z \cup \{i\} \subseteq X$*, we have*

$$PrefixUtil_{S_X}(Z \cup \{i\}) = PrefixUtil_{S_X}(X).$$

Clearly, $PrefixUtil_{S_X}(X) = PrefixUtil_D(X)$*. Thus,*

$$PrefixUtil_D(Z \cup \{i\}) = PrefixUtil_D(X) + PrefixUtil_{D-S_X}(Z \cup \{i\}).$$

Since $PrefixUtil_{D-S_X}(Z \cup \{i\}) \geq 0$*,*

$$PrefixUtil_D(Z \cup \{i\}) \geq PrefixUtil_D(X).$$

This lemma means that the prefix utility of an itemset X has the *download closure property* if we only concern the subsets of X that contain the last item in X in the ranked order. Such a special kind of the downward closure property allows us to use $PrefixUtil$ to prune search space in our HUI mining algorithm to be described later.

Example 3. *Assume that a, b and c are items in a data set and that the items in the data set are ranked in the lexicographic order. According to Lemma 1,* $PrefixUtil(\{ac\}) \geq PrefixUtil(\{abc\})$ *and* $PrefixUtil(\{bc\}) \geq PrefixUtil(\{abc\})$*. Thus, if* $PrefixUtil(\{ac\})$ *or* $PrefixUtil(\{bc\})$ *is less than a minimum utility threshold,* $PrefixUtil(\{abc\})$ *must be less than the threshold. Since* $PrefixUtil(\{abc\}) \geq u(\{abc\})$*,* $u(\{abc\})$ *must be less than the threshold.*

The second challenge of our problem is in finding top-k patterns. An efficient method for finding top-k patterns is to first find potential patterns whose (estimated) utility is above a threshold and then identify the top-k patterns from the potential ones. Since the minimum utility threshold is not given in the top-k problem, a challenge in top-k pattern mining is how to set up the threshold so that the process generates fewer number of potential patterns that include all the top-k patterns. To meet this challenge, we propose some strategies for initializing and dynamically raising the minimum utility threshold during the stream mining process. Below we define *minimum transaction utility*, which will be used in our strategy for initializing the threshold.

Definition 9. ***Minimum Transaction Utility (mtu)*** *of a transaction* T *is defined as:* $mtu(T) = \min_{i \in T}(u(i, T))$.

For example, in Figure 2:

$$\begin{aligned} mtu(T_4) &= \min(u(b, T_4), u(c, T_4), u(d, T_4), u(e, T_4)) \\ &= \min(24, 15, 24, 8) = 8 \end{aligned}$$

Based on the mtu values of the transactions, we define an underestimate utility of an itemset in a data set as follows.

Definition 10. ***Minimum Transaction Utility (MTU)*** *of an itemset* X *over a data set* D *is defined as:* $MTU_D(X) = \sum_{X \subseteq T \bigwedge T \in D} mtu(T)$.

Table 1: Summary of Notations

Concept	Description
$u(i,T)$	Utility of item i in transaction T
$u(X)$	Utility of itemset X in a data set
$TWU(X)$	Transaction-Weighted Utility (an overestimated utility)
HUI	High Utility Itemsets
$PrefixUtil(X)$	Prefix Utility of itemset X
$mtu(T)$	Minimum Transaction Utility of transaction T
$MTU(X)$	Minimum Transaction Utility of Itemset X (an underestimated utility)
$LPI(X)$	Lowest Profit Item Utility of Itemset X (an underestimated utility)
$miu(i)$	Minimum Item Utility of item i in any transaction of a data set
$MIU(X)$	Minimum Itemset Utility of Itemset X (an underestimated utility)
$maxUtilList$	List of maximum values of MTUs and LPIs for each level of *HUDS-tree*
$MIUList$	List of top-k MIU values in potential HUIs
$minTopKUtil_i$	Minimum Top-k Utility of the ith sliding window
$PTKHUI$	Potential Top-k High Utility Itemset
$PTKSet$	Set of Potential Top-k High Utility Itemsets

We use $MTU(X)$ to denote $MTU_D(X)$ when the data set D is clear in the text. For example, for the data set in Figure 2:

$$\begin{aligned} MTU(\{bc\}) &= mtu(T_3) + mtu(T_4) + mtu(T_5) \\ &= 3 + 8 + 4 = 15 \end{aligned}$$

Property 2. *For any itemset X in a data set D, the following holds: $MTU_D(X) \leq u_D(X)$.*

Lemma 2. *The minimum transaction utility of an itemset satisfies the downward closure property. That is, for all $Y \subseteq X$, $MTU(Y) \geq MTU(X)$.*

Proof 2. *Since all the transactions containing an itemset X also contains any subset Y of X, $MTU(Y) \geq MTU(X)$.*

The third challenge for mining top-k HUI in streaming data is that there can be a huge amount of data in a data stream. Thus, use of compact memory data structures is necessary in the mining process. To meet this challenge, a compact data structure is used in our method, which can be built with one scan of data. Finding potential patterns is based on the information in this

data structure. This data structure and our method for finding top-k HUIs are described in the next section. For convenience, Table 1 summarizes the concepts and notations we define in this paper.

4. T-HUDS: Top-k High Utility Itemset Mining over Data Stream

In this section, we propose an efficient method (called *T-HUDS*) to find top-k *HUIs* in data streams without specifying a minimum utility threshold. *T-HUDS* works based on a prefix tree, called *HUDS-tree* (High Utility Data Stream Tree), and two auxiliary lists of utility values. *HUDS-tree* dynamically maintains a compressed version of the transactions in a sliding window. The two auxiliary lists each maintain a utility list of length $log_2(k+1)$ or k, where k is the number of top-k itemsets to be returned, and are used to dynamically adjust the minimum utility threshold during the mining process.

4.1. An Overview of T-HUDS Method

The *T-HUDS* method includes three main steps: (1) *HUDS-tree* construction: construct a *HUDS-tree* and two auxiliary lists on a batch of transactions; (2) *HUDS-tree* mining: discover top-k *HUIs* from the current sliding window; and (3) *HUDS-tree* update: once a new batch arrives, inserts the transactions in the new batch into the tree, remove transactions in the oldest batch from the tree if the sliding window had been filled up, and updates the two auxiliary lists.

Algorithm 1 presents an overview of the proposed method. We assume that the data stream comes in batches. Given a batch B_i of transactions, k and the sliding window size ($winSize$), if a *HUDS-tree* does not exist yet (i.e., the batch is the very first one), a *HUDS-tree* is constructed based on the transactions in B_i, and two auxiliary lists, $maxUtilList$ and $MIUList$, are also computed or initialized. If a *HUDS-tree* already exists, the tree and the two auxiliary lists are updated to reflect the addition or changes of transactions in the sliding window. After that, *T-HUDS* calls Algorithm 3 to find top-k HUIs for the new sliding window.

Below we first describe how the *HUDS-tree* is structured and constructed. Then we present our methods for estimating the minimum utility threshold, our top-k HUI mining algorithm and finally our procedure for updating the *HUDS-tree*.

Algorithm 1 T-HUDS

Input: B_i, k, $winSize$, *HUDS-tree*

Output: *Top-k HUIs*

1: **if** HUDS-tree is empty (i.e., B_i is the very first batch B_1) **then**
2: $minTopKUtil_0 \leftarrow 0$
3: Construct a *HUDS-tree* based on B_i (i.e., B_1)
4: Construct the auxiliary list $maxUtilList$ based on the information in the *HUDS-tree*
5: Initialize the auxiliary list $MIUList$ using the top-k miu values of the items (to be defined in later)
6: **else**
7: Call Algorithm 5 to update *HUDS-tree*, $maxUtilList$ and $MIUList$ using B_i and $winSize$
8: Call Algorithm 3 to compute top-k HUIs on the current sliding window with the *HUDS-tree*, $maxUtilList$, $MIUList$ and $minTopKUtil_{i-1}$
9: **return** Top-k HUIs

4.2. HUDS-tree Structure and Construction

The structure of *HUDS-tree* is similar to that of *FP-tree* [12], *UP-tree* [25] or *HUS-tree* [2]. These trees are used to compress a transaction database into a tree. A non-root node in the trees represents an item in the transaction database, and a path from the root to a node compresses the transactions that contains the items on the path. Since the *FP-tree* is used to find frequent itemsets, a node in an *FP-tree* mainly stores the frequency of an itemset represented by the path from the root to the node. The *UP-tree* is for finding high utility itemsets, and thus its node contains not only frequency but also an estimated utility (i.e., TWU) of the itemset. The *HUS-tree* is used for mining high utility patterns over data streams. Thus, its node stores the TWU value of the itemset for each batch in a sliding window to facilitate the update process. Since we are dealing with data streams as well, our *HUDS-tree* is similar to a *HUS-tree*. But instead of storing TWU values, a node in a *HUDS-tree* stores the $PrefixUtil$ of the represented itemset for each batch, which is, as discussed earlier, a closer estimate of the true utility of the itemset than TWU. In addition, to effectively estimate the minimum utility threshold, a node in *HUDS-tree* also stores the MTU value of the itemset for each batch. The node structure of the *HUDS-tree* is described below.

A non-root node in a *HUDS-tree* contains the following fields: *nodeName*, *nodeCounts*, *nodePUtils*, *nodeMTUs* and *succ*. *nodeName* is the name of the item represented by the node. The *nodeCounts* field is an array with $winSize$ elements, where $winSize$ is the number of batches in the sliding window. Each element in *nodeCounts* corresponds to a batch in the current sliding window and registers the number of the transactions in the batch falling onto the path from the root to the node. Let X be the itemset represented by the path. The *nodePUtils* field is an array of $winSize$ elements, each corresponding to a batch and storing the prefix utility of X in the transactions of the batch falling onto the path. Similarly, *nodeMTUs* is an array of the *minimum transaction utilities* (MTU) of X in the transactions falling onto the path for all the batches of the sliding window. Keeping separate information for each batch facilitates the update process, that is, when a new batch B_i arrives, if the oldest batch needs to be removed, it is easy to remove the information of the oldest batch and include the information for the new batch. Finally, *succ* points to the next node of the tree having the same *nodeName*.

Example 4. *A* HUDS-tree*, built from the transactions in sliding window* $SW_1 = \{B_1, B_2\}$ *in Figure 1, is illustrated in Figure 3, where the* $winSize$ *is 2 and thus* $nodeCounts$*,* $nodePUtils$ *and* $nodeMTUs$ *each contains two values. For example, in node* $\langle b : [0,1],[0,15],[0,3]\rangle$*,* $nodeName$ *is b,* $nodeCounts$ *holds [0,1], meaning the number of transactions matching path* $a \rightarrow b$ *is* 0 *in* B_1 *and* 1 *in* B_2*, respectively, and [0,15] and [0,3] are the contents of* $nodePUtils$ *and* $nodeMTUs$*, respectively. Since b appears only in the second batch, its values for* $nodeCounts$*,* $nodeUtils$ *and* $nodeMTUs$ *in the first batch are* 0*. The field succ is not illustrated for the clarity reason.*

Each item has an entry in the header table of the *HUDS-tree*. An entry in the header table contains the name of the item, the $PrefixUtil$ value of the item in the transactions represented by the tree and a *link* pointing to the first node in the *HUDS-tree* carrying the item. The $PrefixUtil$ value of an item is computed by adding up all the *nodePUtils* values of the nodes labeled with the item in the tree.

Given the first batch B_1 of transactions, a *HUDS-tree* is constructed as follows. For each transaction in B_1, we first order the items in the transaction in an order (such as the lexicographic order or the descending external item utility order), and then insert the items into the HUDS-tree in the way

Algorithm 2 Insert Transaction into HUDS-tree

Input: Transaction T, $rootNode$, idx, $batchNumber$

Output: Updated *HUDS-tree*, maxUtilList

let $item_{idx}$ be the idxth item in T
if $\exists$ $node \in$ the children of the $rootNode$ & $nodeName(node) = item_{idx}$ **then**
    $node.nodeUtils[batchNumber] += \sum_{j=1}^{idx} u(item_j, T)$
    $node.nodeCounts[batchNumber]++$
    $node.nodeMTUs[batchNumber] += MTU(T)$
else
    $node.nodeName \leftarrow item_{idx}$
    $node.nodeUtils[batchNumber] \leftarrow \sum_{j=1}^{idx} u(item_j, T)$
    $node.nodeCounts[batchNumber] \leftarrow 1$
    $node.nodeMTUs[batchNumber] \leftarrow MTU(T)$
    add $node$ as a child node of $rootNode$
update the idxth element, $maxUtil_{idx}$, in the $maxUtilList$
if $idx \neq$ the length of T **then**
    $Algorithm2(T, node, idx + 1, batchNumber)$
HUDS$-Tree \leftarrow rootNode$
return HUDS$-Tree$, $maxUtilList$

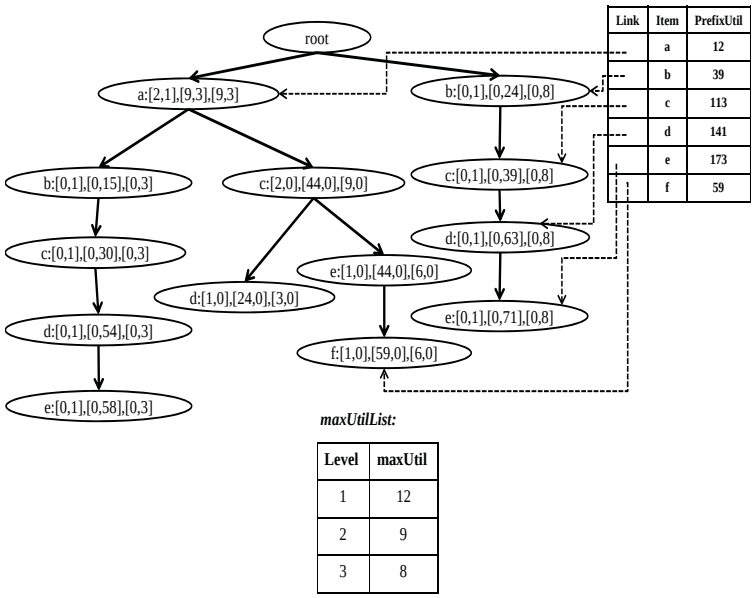


Figure 3: HUDS-tree after inserting transaction in SW_1 in Figure 1.

similar to building an FP-tree [12]. For example, for the first item $item_1$ in a transaction T in B_1, if a node with the same item name is not found under the root, a new child is created and its fields are initialized as follows: $nodeName = item_1$, $nodePUtils[1] = u(item_1, T)$, $nodeCounts[1] = 1$, $nodeMTUs[1] = MTU(T)$. If the node with the item name already exists under the root, its fields for the current batch are updated. Details of the procedure for inserting one transaction T in batch B_i into the *HUDS-tree* are presented in Algorithm 2. In the algorithm, the input parameter $batchNumber$ should be given a value of $i\%winSize + 1$, where i is the ID of the current batch B_i in the data stream and % is the modulo operator which returns the remainder of dividing i by $winSize$. For example, if $i = 2$ or $winSize + 2$, $batchNumber$ is 2. The algorithm is a recursive algorithm. Each call to the algorithm "inserts" one item of the input transaction T into the tree. The input parameter idx indicates which item in T is being "inserted". idx is initialized to 1 for each transaction. Clearly, the tree can be built with one scan of the data in B_i.

Before we describe how to mine HUIs from a *HUDS-tree* and how to update the tree with new batches, we first present our method for estimating the minimum utility threshold.

4.3. Estimation of Minimum Utility Threshold

Our objective is to find top-k high utility itemsets. Since the number of itemsets is exponential with respect to the number of items in the data, it is infeasible to enumerate all the itemsets, find their utilities in the sliding window and outputs the top-k highest utility itemsets. An efficient procedure for finding top-k itemsets is to first use an efficient method to find potential itemsets whose utility is above a threshold and then identify the top-k itemsets from the potential ones. To do this, a proper minimum utility threshold is needed in the first phase of the procedure. If the threshold is set too low, many unwanted HUIs are produced, which is time-consuming. If it is set too high, we may not be able to produce k itemsets. A good strategy for setting the threshold should satisfy the following conditions: (1) it should not miss any top-k *HUIs*; (2) the estimated threshold should be as close as possible to the utility of the kth highest utility itemset.

In our method, we use four strategies to initialize and dynamically adjust the threshold during the mining process. These strategies lead to significant pruning of search space. Below we describe three strategies, which will be used in the first phase of our mining method. The fourth strategy (to be used in the second phase) will be described in Section 4.4.2.

4.3.1. Initializing the Threshold Using maxUtilList

In a *HUDS-tree*, the $nodeMTUs$ field of a node n stores the MTU values of the itemset represented by the path from the root to n in the set of transactions falling onto the path in each batch separately. The MTU value of the itemset in the transactions on the path in the sliding window can be easily calculated by summing up all the values in $nodeMTUs$ of node n. We use $nodeMTU(n)$ to denote this sum. Similarly, $nodeCount(n)$ is used to denote the count of the itemset in the set of the transactions falling on the path in the whole sliding window. Now we are ready to define the $maxUtilList$.

Definition 11. ***(Maximum Utility List (maxUtilList))*** *of a* HUDS-tree *is a list of length d:*

$$maxUtilList = \{maxUtil_1, ..., maxUtil_d\}$$

where d is the depth of the HUDS-tree *and* $maxUtil_i$ *is computed based on*

the nodes on the ith level of the tree as follows:

$$\begin{aligned} maxUtil_i &= \max_j\{\max(nodeMTU(node_{i,j}), \\ &\quad nodeCount(node_{i,j}) \times minProfit(node_{i,j}))\} \end{aligned}$$

where $node_{i,j}$ is the jth node in level i of the tree, $nodeMTU(node_{i,j})$ is sum of the values in the $nodeMTUs$ field of $node_{i,j}$ (i.e., the total MTU value in the sliding window), $nodeCount(node_{i,j})$ is the sum of the counts in the $nodeCounts$ field of $node_{i,j}$ (i.e., the total count in the sliding window), and $minProfit(node_{i,j}) = \min\{p(item)|item \in X\}$ where $p(item)$ is the external utility of the item and itemset X is formed by the path from the root to $node_{i,j}$ in the tree.

For example, assume that the root is at level 0 in Figure 3. The level 2 has one b node and two c nodes. $maxUtil_2$ is thus computed as:

$$\begin{aligned} maxUtil_2 &= \max\{\max(nodeMTU(b), nodeCount(b) \times 3), \\ &\qquad \max(nodeMTU(c), nodeCount(c) \times 3), \\ &\qquad \max(nodeMTU(c), nodeCount(c) \times 5)\} \\ &= \max\{\max(3, 1 \times 3), \max(9, 2 \times 3), \max(8, 1 \times 5)\} = 9. \end{aligned}$$

Lemma 3. *Let $util_k$ be the utility of the kth itemset in the top-k high utility itemset list. $util_k$ is no less than $maxUtil_L$ where $L = \lceil \log_2(k+1) \rceil$.*

Proof 3. *Let's call $nodeCount(node_{i,j}) \times minProfit(node_{i,j})$* Lowest Profit Item utility (LPI) *of the itemset X formed by the path from the root to $node_{i,j}$ in the set S of transactions represented by the path. Clearly, $LPI(X)$ is another underestimate of the utility of X in S, i.e., $LPI(X) \leq u(X)$ on S. Also, for all $Y \subseteq X$, $LPI(Y) \geq LPI(X)$ on S.*

Let $node_{L,j}$ be a node on level L of the tree, $X_{L,j}$ denote the itemset formed by the path from the root to $node_{L,j}$, and $S_{L,j}$ denote the set of transactions falling onto the path. Assume that $node_{L,j}$ is the node with $maxUtil_L$, that is, $maxUtil_L$ is either $nodeMTU(node_{L,j})$ (i.e., $MTU(X_{L,j})$ on $S_{L,j}$) or $LPI(X_{L,j})$ on $S_{L,j}$.

Assume that Y is a subset $X_{L,j}$. According to Lemma 2, $MTU(Y) \geq MTU(X_{L,j})$ on set $S_{L,j}$. According to Property 2, $u(Y) \geq MTU(Y)$ on $S_{L,j}$. Similarly, $u(Y) \geq LPI(Y) \geq LPI(X_{L,j})$ on $S_{L,j}$. Thus,

$$u(Y) \geq max(nodeMTU(node_{L,j}), LPI(X_{L,j})) = maxUtil_L.$$

Since $u(Y)$ on the entire data set represented by the tree is no less than $u(Y)$ on $S_{L,j}$. Thus, $u(Y)$ on the entire data set is no less than $maxUtil_L$.

Since $node_{L,j}$ is at level L of the tree, $X_{L,j}$ contains L items (assuming the root is at level 0). Thus, $X_{L,j}$ has $2^L - 1$ subsets. Thus, there are at least $2^L - 1$ itemsets whose utility is no less than $maxUtil_L$.

If $L = \lceil \log_2(k+1) \rceil$, we have

$$L \geq \log_2(k+1) \Rightarrow 2^L \geq k+1 \Rightarrow 2^L - 1 \geq k$$

Thus, there are at least k itemsets with utility higher than or equal to $maxUtil_L$. Thus, $util_k$ is no less than $maxUtil_L$.

Lemma 3 declares that $maxUtil_L$ can be used to set the minimum utility threshold for finding top-k HUIs, where $L = \lceil \log_2(k+1) \rceil$. No top-$k$ HUIs can be missed with such a threshold. Intuitively, $maxUtil_L$ is the maximum value among the $nodeMTU$ values and LPI values of the nodes on level L of the tree.

The *maxUtilList* can be computed while constructing and updating the *HUDS-tree*. If k is fixed, only $maxUtil_L$ needs to be computed in the list; otherwise, the values of $maxUtil_i$ for all the levels are maintained.

4.3.2. Adjusting the Threshold Using $MIUList$

$MIUList$ is another list that we maintain to dynamically adjust the minimum utility threshold. It keeps the top-k *minimum itemset utility* (MIU) values of current potential high utility itemsets. Below we first define the concept of MIU [27]:

Definition 12. ***Minimum Item Utility of an item*** *a* ***in any transaction of a dataset*** *D is defined as: $miu_D(a) = u(a, T_q)$ where $T_q \in D$ and $\neg\exists\ T_p \in D$ such that $u(a, T_p) < u(a, T_q)$*

Definition 13. ***Minimum Itemset Utility of an itemset*** *X in a dataset D is defined as:*
$MIU_D(X) = \sum_{a_i \in X} miu_D(a_i) \times SC_D(X)$ where $SC_D(X)$ is support count of X in D.

We use $MIU(X)$ to denote $MIU_D(X)$ when the data set D is clear in the context.

Property 3. *For any itemset X in dataset D, $MIU_D(X) \leq u_D(X)$.*

The *miu* value of an item can be computed during the *HUDS-tree* construction and update. It can be stored in the global header table of the *HUDS-tree*. The MIU value of an itemset can be computed based on the *miu* values of its elements and the support count of the itemset (maintained in the *nodeCounts* fields). In [27], the MIU values of itemsets are used to raise the minimum support threshold during the HUI mining process. But they may not be used properly. We use them to adjust the minimum utility threshold by maintaining a *minimum itemset utility list* defined as follows.

Definition 14. ***Minimum Itemset Utility List(MIUList)*** Given a set of already-generated HUIs, $MIUList$ contains the top-k list of the MIU values of these HUIs, ranked in MIU-descending order, denoted as $MIUList = \{MIU_1, MIU_2, ..., MIU_k\}$, where $MIU_1 \geq MIU_2 \cdots \geq MIU_k$.

Lemma 4. *Let MIU_k be the* kth *member of* MIUList *and $util_k$ be the utility of the kth highest utility itemset in the top-k HUI list. $util_k$ is no less than MIU_k.*

Proof 4. *Assume that the MIU_i values in $MIUList$ are the MIU values of itemsets X_1, X_2, $\ldots X_k$, respectively. According to Property 3, we have:*

$$\forall X_i \in \{X_1, X_2, \ldots X_k\}, MIU(X_i) \leq u(X_i).$$

According to the Definition 14, MIU_k is the smallest value in the $MIUlist$. Thus, there are at least k itemsets whose utility is no less than MIU_k.

According to this lemma, if the minimum utility threshold is set to MIU_k, no top-k HUI will be missed. Thus, we have the following strategy for adjusting the threshold. Once the *HUDS-tree* is built or updated for a sliding window SW_i, $MIUList$ is initialized to the top-k highest *miu* values of single items. During the process of mining HUIs for window SW_i, once a new potential HUI is generated, its MIU is compared with the current MIU_k. If it is greater than the current MIU_k, the new MIU value is inserted into the $MIUList$. If the new MIU_k is greater than the current minimum utility threshold, then the threshold can be raised to the new MIU_k.

4.3.3. Adjusting the Threshold with minTopKUtil of Last Window

Our third strategy for adjusting the minimum utility threshold is to make use of the utility values of the top-k HUIs in the last sliding window. For this, we define the *minimum top-k utility* ($minTopKUtil$) of a sliding window as follows.

Definition 15. *Let $SW_i = \{B_i, B_{i+1}, \ldots, B_{i+winSize-1}\}$ be the ith sliding window and let $TopkHUISet_i$ denote the set of top-k HUIs in window SW_i. The minimum top-k utility of a sliding window SW_i is defined as:*

$$minTopKUtil_i = \min_{itemset \in TopkHUISet_i} \sum_{j=i+1}^{i+winSize-1} u_{B_j}(itemset)$$

In other words, the $minTopKUtil$ of sliding window SW_i is the minimum of the utilities of the itemsets in $TopkHUISet_i$ in the last $winSize - 1$ batches of SW_i.

Lemma 5. *Let $util_k$ be the utility of the kth highest utility itemset over sliding window SW_{i+1}, and $minTopKUtil_i$ be the minimum top-k utility of window SW_i. We have $util_k \geq minTopKUtil_i$.*

Proof 5. *Let B be the union of last $winSize - 1$ batches in window SW_i. Then the next sliding window $SW_{i+1} = B \cup B_{new}$ where B_{new} is the new batch in SW_{i+1}. Since $B \subset SW_{i+1}$, for each itemset X in $TopkHUISet_i$, $u_B(X) \leq u_{SW_{i+1}}(X)$. Since $minTopKUtil_i \leq u_B(X)$ for all $X \in TopkHUISet_i$ and there are k itemsets in $TopkHUISet_i$, there are at least k itemsets whose utility in SW_{i+1} is at least $minTopKUtil_i$.*

According to this lemma, if the minimum utility threshold in window SW_{i+1} is set to $minTopKUtil_i$, no top-k high utility itemsets will be missed.

The $minTopKUtil_i$ value is computed during the second phase of our procedure for mining top-k HUIs from sliding window SW_i, which is to be described in Section 4.4.2.

4.4. Mining Top-k High Utility Itemsets

After a *HUDS-tree* is built or updated for a sliding window SW_i, we use a 2-phase procedure to find top-k HUIs in SW_i. In the first phase, the *HUDS-tree* is mined to generate a set of potential top-k high utility itemsets (i.e.,

Algorithm 3 Top-k HUI Mining

Input: $HUDS\text{-}Tree$, $maxUtilList$,$MIUList$, $minTopKUtil_{i-1}$, k, SW_i
Output: $TopkHUISet$, $minTopKUtil_i$

1: $L \leftarrow \lceil \log(k+1) \rceil$
2: $min\_util \leftarrow \max\{maxUtil_L, MIU_k, minTopKUtil_{i-1}\}$
3: Generate a set of potential top-k HUIs (PTKSet) by calling Algorithm 4 with min_util. The min_util is also dynamically updated in Algorithm 4
4: Scan the transactions in the current sliding window SW_i to obtain $u_{SW_i}(itemset)$ and $u_{SW_i - B_i}(itemset)$ for each $itemset$ in $PTKSet$, where B_i is the first batch in SW_i.
5: $TopkHUISet \leftarrow \emptyset$
6: **for** each $itemSet \in PTKSet$ **do**
7: **if** $u_{SW_i}(itemSet) \geq min\_util$ **then**
8: Insert $\langle itemSet, u_{SW_i}(itemSet) \rangle$ into $TopkHUISet$ so that the elements in $TopkHUISet$ are ranked in the utility-descending order
9: **if** the size of $TopkHUISet > k$ **then**
10: Remove the last element from $TopkHUISet$
11: **if** $u_{SW_i}(lastItemSet) > min\_util$ where $lastItemSet$ is the current last itemset in $TopkHUISet$ **then**
12: $min\_util \leftarrow u_{SW_i}(lastItemSet)$
13: $minTopKUtil_i \leftarrow \min\{u_{SW_i - B_i}(itemset) | itemset \in TopkHUISet\}$
14: **return** $TopkHUISet$, $minTopKUtil_i$

$PTKHUI$s) that satisfy a dynamically-changing minimum utility threshold. In the second phase, the exact utilities of the $PTKHUI$s are computed and the top-k high utility itemsets are returned.

This 2-phase procedure is shown in Algorithm 3. At the beginning of the procedure, we initialize the minimum utility threshold, *min_util*, according to the strategies proposed in Section 4.3 as follows:

$$min\_util = \max\{minTopKUitl_{i-1}, maxUtil_L, MIU_k\}.$$

where $minTopKUitl_{i-1}$ is the minimum top-k utility of the last sliding window (initialized to 0 in Algorithm 1 if the new batch is the first one), $maxUtil_L$ is the Lth element in *maxUtilList* (where L is computed in Line 1), and MIU_k is the kth element of the $MIUList$ that initially contains the list of the top-k minimum item utilities (*miu*) of single items.

With this initial *min_util* threshold, Algorithm 4 is called to find PTKHUIs from the *HUDS-tree* (Line 3). This is the first phase of the top-k procedure. The second phase (from Line 4 to the end) finds exact top-k HUIs from the set of PTKHUIs. Below we describe each phase in detail.

4.4.1. Phase I: Discover $PTKHUI$s from HUDS-tree

In Phase I, a set of potential top-k HUIs ($PTKHUI$s) is found from the *HUDS-tree*. Our objective in this phase is to find as few $PTKHUI$s as possible (so that the second phase will be faster) while not missing any top-k HUIs. Our procedure for this phase follows a pattern growth approach, similar to *FP-growth* [12] and $HUPMS$ [2]. The major differences between our Phase I procedure and the others are as follows. First, we use both $PrefixUtil$ and local TWUs to prune the search space, while others for HUI mining mainly use TWU. Second, we use effective strategies for initializing and dynamically adjusting the *min_util* threshold during the mining process.

The pseudocode of the *HUDS-tree* mining procedure is described in Algorithm 4. Like *FP-growth*, the algorithm is a recursive algorithm. In the first call to the procedure, the input *HUDS-tree* is the global tree, and the itemset X in the input list is empty. In a recursive call, the input tree is the X-conditional *HUDS-tree* where X is a non-empty itemset. The algorithm works as follows. For each item t in the (conditional) header table, the algorithm checks if the $PrefixUtil$ of t satisfies the *min_util* threshold (Line 2). If yes, a potential top-k HUI IS is generated by extending X with item t. IS is then added into the potential top-k HUI set (i.e., *PTKSet*). Then,

Algorithm 4 HUDS-tree Mining to Generate PTKHUIs (Phase I)

Input: $HUDS\text{-}Tree$, itemset X, min_util, $MIUList$, k
Output: $PTKSet$, min_util,$MIUList$

1: **for** each item t in the header table of *HUDS-tree* **do**
2: **if** $PrefixUtil(t) \geq min\_util$ **then**
3: Generate a potential top-k itemset: $IS \leftarrow \{t\} \cup X$
4: Add IS into the $PTKSet$ set
5: **if** $MIU_{SW_i}(IS) \geq min\_util$ **then**
6: Insert $MIU_{SW_i}(IS)$ into the $MIUList$
7: $min\_util \leftarrow MIU_k$
8: $Pattern\_base_{IS} \leftarrow$ all prefix paths of the nodes for item t with their utilities
9: Prune all items in the $Pattern\_base_{IS}$ whose TWU in $Pattern\_base_{IS}$ is less than min_util.
10: Construct conditional HUDS$-$Tree$_{IS}$ and its header table
11: **if** $HUDS\text{-}Tree_{IS}$ is not empty **then**
12: call Algorithm 4($HUDS\text{-}Tree_{IS}$,IS, min_util, $MIUList$, k)
13: **return** $PTKSet$, min_util,$MIUList$

the *min_util* threshold is adjusted in lines 5 to 7. If *MIU(IS)* is more than the current min_util, the MIU value is inserted into $MIUList$ and min_util is raised by the minimum value of $MIUList$. $MIU(IS)$ can be computed easily because $SC_{SW_i}(IS)$ can be computed using the *nodeCounts* fields of the t nodes and the *miu* values of all the items have already been computed when building the global *HUDS-tree*.

After IS is generated, to find longer PTKHUIs containing IS, IS-conditional pattern base ($Pattern\_base_{IS}$) is built by enumerating all the prefix paths of the t nodes in the tree. The utility of each prefix path is the sum of the values in the *nodePUtils* field of the t node in that path. Each item's local TWU value can then be computed by adding up the utilities of the prefix paths it is in. In Line 9, we eliminate items in the conditional pattern base whose local TWU is less than the min_util threshold. After that, the IS-conditional *HUDS-tree* is constructed based on the conditional pattern base with the remaining items. At the end of tree construction, all the *nodePUtils* values of nodes with the same *nodeName* in the conditional tree are added and the result is added to local header table as the $PrefixUtil$ value of the item. Once a conditional tree is built, Algorithm 4 is called recursively to

discover longer $PTKHUI$s ending with IS.

In the performance evaluation section, we will show that this pattern-growth procedure generates fewer potential top-k HUIs and has less overall run time than the state-of-the-art algorithms for high utility itemset mining. This is due to the use of the prefix utility in pruning the search space and also the dynamical increase of *min_util* during the mining process.

4.4.2. Phase II: Identifying Top-k HUIs from PTKHUIs

HUDS-tree is a compact representation of the transactions in a sliding window. It allows the use of the pattern growth method to efficiently find the potential top-k HUIs. However, since the quantity of an item inside a transaction may vary among transactions, the exact utility of an itemset cannot be obtained from the *HUDS-tree*. Thus, in this second phase, we scan the transactions in the current sliding window to obtain the exact utility of each potential top-k HUI, and then identify the top-k HUIs based on the true utility of the PTKHUIs.

The second phase procedure is shown in Lines 4-12 of Algorithm 3. In Line 4, it scans the transactions in the current sliding window SW_i to obtain the exact utility of each itemset in $PTKSet$ in SW_i and also the exact utility of each itemset in the last $winSize-1$ batches of SW_i. From Line 6 to Line 12, top-k HUIs are identified using a *selected insertion sort*, in which only the itemsets whose utility is no less than *min_utility* are inserted to the top-k list (denoted as $TopkHUISet$). $TopkHUISet$ is maintained to have no more than k elements, ranked in utility-descending order. In addition, if $TopkHUISet$ contains k elements, *min_util* is adjusted dynamically to be the utility of the kth itemset in $TopkHUISet$ (Lines 11 and 12). We call this adjustment our fourth strategy for increasing the *min_util* threshold.

Finally, in Line 13 of the algorithm, the minimum top-k utility of the current sliding window (SW_i) is set to minimum utility value of the itemset in $TopkHUISet$ in the last $winSize-1$ batches of SW_i. This is for adjusting the *min_util* threshold for mining tip-k HUIs in the next sliding window SW_{i+1}.

Theorem 1. *Given a sliding window SW_i, if X is among the top-k high utility itemsets, it is returned by Algorithm 1.*

Proof 6. *We prove the theorem by showing that the min_util in our algorithm is never over the exact utility of the kth highest utility itemset in the*

current sliding window, and also that our HUDS-tree *mining procedure does not prune out any itemset whose true utility is greater than min_util.*

Let $util_k$ be the exact utility of the kth highest utility itemset for sliding window SW_i. In our algorithms, the min_util is set or adjusted in the following three places:

- *In Line 2 of Algorithm 3:*

$$min\_util = \max\{maxUtil_L, MIU_k, minTopKUtil_{i-1}\}$$

  where $L = \lceil \log_2(k+1) \rceil$. According to Lemmas 3, 4, and 5, $maxUtil_L \leq util_k$, $MIU_k \leq util_k$ and $minTopKUtil_{i-1} \leq util_k$. Thus, min_util is no larger than $util_k$.
- *In Lines 5-7 of Algorithm 4, min_util is dynamically adjusted to MIU_k, which is the kth highest MIU value of the already generated potential top-k HUIs. According to Lemma 4, $MIU_k \leq util_k$. Thus, $min\_util \leq util_k$.*
- *In Lines 11-12 in Algorithm 3, min_util is dynamically adjusted to the lowest utility of the current top-k HUI set. Thus, min_util is no larger than $util_k$.*

Below we show that our HUDS-tree *mining procedure for generating potential top-k HUIs (i.e., Algorithm 4) does not miss any top-k HUIs. There are two places where we prune the search space in Algorithm 4.*

- *In Line 2, if the PrefixUtil of an item t is less than min_util, item t will not be added to itemset X to form longer HUI containing $\{t\} \cup X$. The $PrefixUtil$ of t in the (conditional) header table is actually $PrefixUtil(\{t\} \cup X)$ (according to how it is computed). Assume $X = Y \cup \{i\}$ where i is the last item in X in the item order for building the* HUDS-tree. *Then $\{t\} \cup X = \{t\} \cup Y \cup \{i\}$. According to Lemma 1, $PrefixUtil(\{t\} \cup Y \cup \{i\}) \geq PrefixUtil(S \cup \{t\} \cup Y \cup \{i\})$ where S is a set of items containing the items ranked before t in the item order for building the tree. Thus, if $PrefixUtil(\{t\} \cup Y \cup \{i\}) < min\_util$, $PrefixUtil(S \cup \{t\} \cup Y \cup \{i\}) < min\_util$. This means that if the $PrefixUtil$ of t in the header table is less than min_util, there is no need to check any itemsets whose "suffix" is $\{t\} \cup X$.*
- *In Line 9 of the algorithm, we prune out all the items whose local TWU is less than min_util. Since TWU has the downward closure property, the pruning does not miss any itemsets whose TWU is no less than min_util*

Both $PrefixUtil$ and TWU are over-estimates of the true utility of an itemset. If an over-estimate is less than min_util, the true utility must be less than min_util. Thus, if an itemset is pruned by $PrefixUtil$ or TWU, its true utility must be less than min_util. Thus, no itemsets whose utility $\geq$ min_util is pruned by the algorithm. Since min_util is never over $util_k$, no top-k HUI is missed by our algorithms.

4.5. HUDS-tree Update

Algorithm 5 HUDS-tree-Update

Input: $HUDS\text{-}Tree$, new batch B_i, k
Output: $HUDS\text{-}Tree$, $maxUtilList$,$MIUList$

1: $batchNumber \leftarrow i\%winSize + 1$
2: **if** the batch ID $i > winSize$ **then**
3: **for** each node in *HUDS-tree* **do**
4: $nodeCounts[batchNumber] \leftarrow 0$
5: $nodePUtils[batchNumber] \leftarrow 0$
6: $nodeMTUs[batchNumber] \leftarrow 0$
7: **if** $\forall i$ $(1 \leq i \leq winSize)$ $nodePUtils[i] = 0$ **then**
8: remove the node and its subtree from the tree
9: **for** every $T \in B_i$ **do**
10: $\{HUDS\text{-}Tree, maxUtilList\} \leftarrow$ Algorithm2(T, *HUDS-Tree*, root of *HUDS-Tree*, 1, $batchNumber$)
11: update the miu value of each item in T
12: Update the $PrefixUtil$ value of each item in the header table by summing up all the values in the $nodePUtils$ fields of all the nodes for the item in the tree.
13: Update $MIUList$ by (1) computing the MIU value of each item in the header table using the miu value of the item and the $nodeCounts$ values in all the nodes for the item and (2) select the top-k MIU values.
14: **return** $HUDS\text{-}Tree$, $maxUtilList$,$MIUList$

When a new batch of transactions arrives, the *HUDS-tree* needs to be updated to represent the transactions in the new sliding window. This involves removing from the tree the information of the oldest batch in the last window (if the last window was full) and adding to the tree the transactions in the new batch. Algorithm 5 describes this update process.

In Line 1, the index of the batch in the tree node fields is computed as $batchNumber = i\%winSize + 1$, where i is the new batch ID (assuming the very first batch in the data stream is B_1), and $winSize$ is the maximum number of batches in a sliding window. The information about the new batch will be put into the $batchNumber$th slots in the $nodeCounts$, $nodePUtils$ and $nodeMTUs$ fields of the tree nodes. In Lines 2 to 8, if the new batch ID (i.e., i in B_i) is greater than the size of the sliding window (which means that the last sliding window was full), then the information about the oldest batch is removed by changing $nodeCounts[batchNumber]$, $nodePUtils[batchNumber]$ and $nodeMTUs[batchNumber]$ in each node to zero. If the sum of the values in $nodePUtils$ for all the remaining batches is zero in a node, the node and the subtree rooted at the node are removed (Line 7). Then, the transactions in the new batch are inserted into the tree one by one by calling Algorithm 2. $batchNumber$ is passed to Algorithm 2 so that the information about the new batch will be stored the $batchNumber$th slots in the node fields. In Algorithm 2, $maxUtilList$ is also updated. After all the transactions are inserted into the tree, the prefix utilities of each item is updated in Line 12. Finally, the $MIUList$ is updated as described in Line 13.

5. Performance Evaluation

In this section, the proposed method for finding top-k high utility itemsets over data stream is evaluated. All the algorithms are implemented in Java. The experiments are conducted on an Intel(R) Core(TM) i7 2.80 GHz computer with 4 GB of RAM.

5.1. Datasets and Performance Measures

Three datasets are used in our experiments. The first one is IBM synthetic dataset *T10I4D100K* [9] where the numbers after T, I, and D represent the average transaction size, average size of maximal potentially frequent patterns, and the number of transactions, respectively. The other two datasets are real life datasets *BMS-POS* and *ChainStore*. *BMS-POS* contains several years worth of point-of-sale data from a large electronics retailer [9]. *ChainStore* is a dataset with over a million transactions, obtained from [22]. Table 2 shows details of the datasets. The *ChainStore* dataset already contains external utilities of the items and the frequency of each item in a transaction. But the two other datasets do not provide external utility or the quantity of

Table 2: Details of the datasets

Dataset	# Trans.	# Items	Avg.Length	batchSize	winSize
IBM	100,000	870	10.1	10,000	5
BMS-POS	515,597	1,657	6.53	50,000	4
ChainStore	1,112,949	46,086	7.2	100,000	6

each item in each transaction. Hence, we randomly generated these numbers using a method described in [2] as follows. The external utility of each item is generated between 1 and 10 by using a *log-normal* distribution and the quantity of each item in a transaction is generated randomly between 1 and 10. $batchSize$ in the Table 2 shows how many transactions are in a batch. It is set in the same way as in [2] so that each data set has around 10 batches. The last column, $winSize$, shows the number of batches in a sliding window. We will later change the $winSize$ setting to show the effect of $winSize$ on performance measures.

We use the following performance measures in our experiments: (1) *number of generated candidates*: the total number of generated PTKHUIs at the end of phase I among all the sliding windows, (2) *Threshold*: the threshold value obtained at the end of method execution, (3) *Run Time (seconds)*: the total execution time of the method over all the sliding windows, (4) *First Phase Time (seconds)*: the total run time of the algorithm for phase I (generating *PTKHUIs*) over all the windows, (5)*Second Phase Time (seconds)*: the total run time of each algorithm for phase II (finding *Top-HUI* set) over all the windows, (6) *Memory Usage(Mega Bytes)*: the memory consumption of the algorithm, average over all the sliding windows.

5.2. Methods in Comparison

To the best of our knowledge, there does not exist a top-k high utility itemset mining method over data streams. Hence, two modified approaches are implemented as comparison methods. The first one is the method proposed in [27] which discovers top-k high utility itemsets from a static data set based on the UP-Growth method [25]. Since this method is not applicable to data streams, we run this method on each sliding window individually, and collect the aggregated values for the performance measures. This method is named TKU. TKU has different versions, each employing a different set of threshold-raising strategies [27]. Here we use their base method plus those threshold-raising strategies whose computation requires only one scan of data

and can be done based on the information from our *HUDS-tree*. TKU sets up its initial threshold to either 0 or the kth highest value of the lower bounds for the utility of certain 2-itemsets. But to get these lower bounds, we need to scan the data set twice to compute them, which is not acceptable for data streams. Thus, we initialize the minimum utility threshold to small value (0.01%) at the beginning of TKU [27].

The second method that we compare our method with is the *HUPMS* algorithm [2], which discovers all the high utility itemsets over data streams given a user-input minimum utility threshold. To compare with the top-k mining methods, we run the HUPMS algorithm with a minimum utility threshold being the threshold raised at the end of the Phase I execution of *TKU*. This is a fair choice of the threshold because a too low threshold would certainly make *HUPMS* very time-consuming, and a too high threshold would unfairly favor *HUPMS* in terms of run time. We denote this *HUPMS* method that uses a threshold from *TKU* as $HUPMS_T$ in our results.

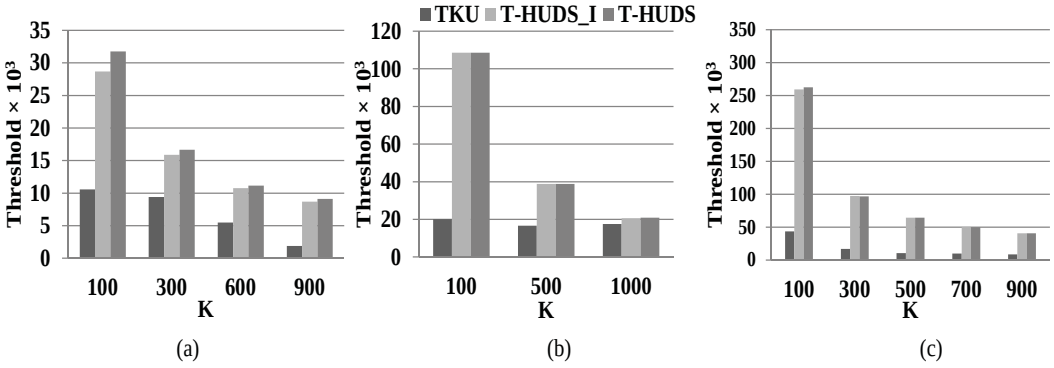
We also compare our method *T-HUDS* with *HUPMS* in terms of HUI mining with different user-specified minimum utility thresholds. In such a comparison, we do not use any threshold-raising strategies in *T-HUDS*, but let it return all the HUIs satisfying the input utility threshold. The purpose of such a comparison is to see the effect of using $PrefixUtil$ to prune the search space in comparison to the use of TWU as in *HUPMS*.

To see how effective our threshold-setting/raising strategies is in the first phase of the method, we use two versions of our T-HUDS method to compare with TKU and $HUPMS$. The first one, denoted as $T\text{-}HUDS_I$, uses only the 3 strategies that apply to the first phase of our method. The second one, denoted as *T-HUDS* is the full version of our method that uses all the 4 strategies, including the one in the second phase.

5.3. Effectiveness of the Obtained Threshold

Figure 4 shows the threshold values obtained at the end of different methods on three datasets. Since *HUPMS* does not raise the threshold during the mining process, we just compare the results of *TKU* with the proposed methods. This figure shows that $T\text{-}HUDS_I$ and *T-HUDS* have similar performance and their final thresholds are much higher than *TKU*. Since none of these three methods miss any top-k HUIs, the higher the final threshold, the better the method. Thus, both T-HUDS$_I$ and *T-HUDS* significantly outperform *TKU*. Between T-HUDS$_I$ and *T-HUDS*, *T-HUDS* is bit better, but not significantly. This means that the 3 strategies used in Phase I of *T-HUDS* are

Figure 4: Reached threshold on (a) IBM, (b) BMS-POS,(c) ChainStore Datasets



very effective, raising the threshold close to the exact utility of the kth highest utility itemset. Recall that the threshold value at the end of Phase II is the exact utility of the kth itemset in the top-k list.

The figure also shows that the threshold value decreases when k increases. It is because the larger the k value is, the lower the threshold value needs to be to return more itemsets. In addition, the figure shows that the difference between *TKU* and our methods is more significant when the size of the data set becomes larger. For example, the difference on *ChainStore* is much bigger than the one on the *IBM* dataset.

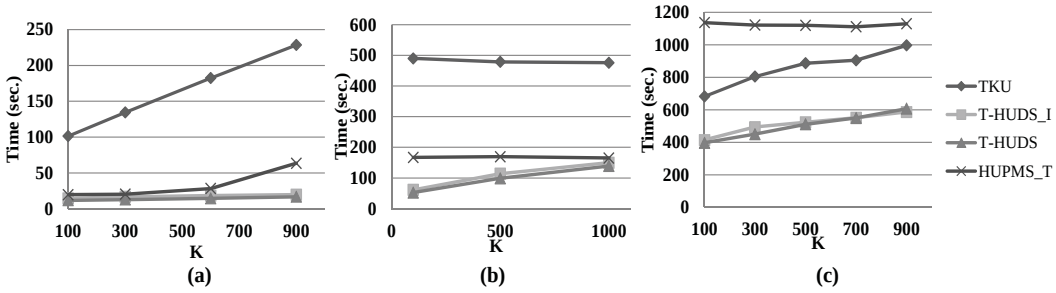
5.4. Number of generated candidates

In addition to the obtained threshold, the number of generated candidates (i.e., *PTKHUI*s) at the end of the first phase is another metric to assess the effectiveness of HUI mining methods. Table 3 presents the numbers of generated candidates on different datasets from different methods for different k values. The numbers show that *T-HUDS* significantly outperforms *TKU*. The results for *T-HUDS*$_I$ are not shown here because they are the same as the ones for *T-HUDS*. The table also shows that *HUPMS*$_T$ method generates fewer candidates in smaller datasets than *T-HUDS*, but much more candidates on larger data sets. The number of candidates generated by *HUPMS*$_T$ is determined by the minimum utility threshold given to the method, which is the threshold reached at the end of Phase I of TKU. Even though the final Phase I threshold of *T-HUDS* is much higher than that of TKU, the number of candidates generated by *HUPMS*$_T$ can still be smaller than that

Table 3: Number of candidates generated in phase I

Dataset	K	TKU	T-HUDS	$HUPMS_T$
IBM	100	2852013	69959	22038
	300	4939423	84898	26668
	600	8177111	94850	54969
	900	10183472	100875	217874
BMS-POS	100	57315	35697	31407
	500	92684	44320	42467
	1000	113842	52195	62512
ChainStore	100	222037	19751	101435
	300	275249	32213	152451
	500	305301	77635	201531
	700	326041	132759	242027
	900	371008	227826	282074

Figure 5: Run time on (a) IBM, (b) BMS-POS, (c) ChainStore Datasets

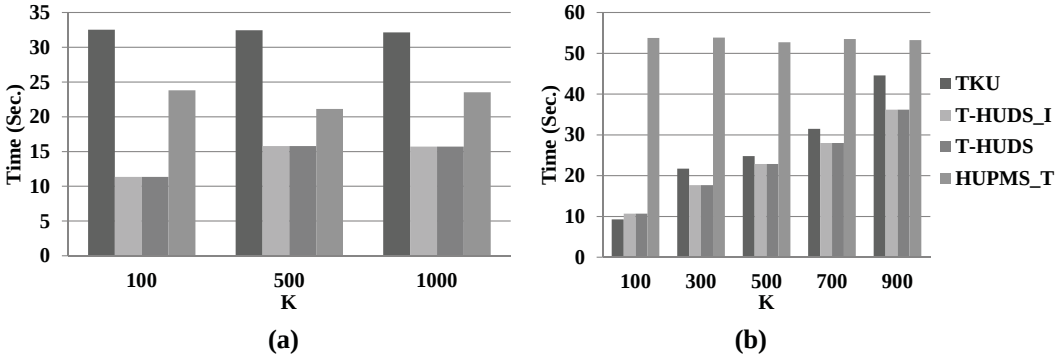


from *T-HUDS*. This is because the initial threshold of *T-HUDS* can be lower than the final Phase I threshold of TKU. But on very large data set (such as *ChainStore*), the initial threshold of *T-HUDS* can be higher than or close to the final Phase I threshold of TKU since the number of candidates generated by $HUPMS_T$ is much higher than the one by *T-HUDS*.

5.5. Efficiency of T-HUDS: Run Time

Figure 5 shows the total run time of each method, including the run time for both Phase I and Phase II. On the IBM and BMS-POS datasets, the execution time of *TKU* algorithm is much worse than others, and $HUPMS_T$ is a bit worse than $T\text{-}HUDS_I$ and *T-HUDS*. But on *ChainStore*, $T\text{-}HUDS_I$ and *T-HUDS* are significantly faster than both $HUPMS_T$ and *TKU*. On this

Figure 6: Run Time for Phase I: (a) BMS-POS, (b) ChainStore



largest data set, $HUPMS_T$ is the worst, even much worse than *TKU*. The run time for *T-HUDS*$_I$ and *T-HUDS* are very similar, although *T-HUDS* is slightly faster due to its raising *min_util* dynamically for pruning out unpromising itemsets in Phase II. Also, it can be observed that the run time of the proposed methods are not affected significantly by the k values, and it increase slightly or slowly when k increases.

To see how each method works in different phases, Figures 6 and 7 present the execution time for Phases I and II, respectively. In Phase I, two proposed methods have the same performance. But in the second phase, *T-HUDS* is more efficient. This is because it dynamically increases the *min_util* threshold in Phase II and consequently the number of candidates compared with the running top-k list is fewer than that in *T-HUDS*$_I$.

5.6. Memory Usage

Since all the algorithms under our comparison need to store the transactions in the current window, we only report the memory usage taken by the trees, their header tables and auxiliary data structures. Table 4 reports the memory consumption on the three datasets. *TKU* consumes the most memory, even though the structure of its tree node is the smallest among the three methods. This is due to the larger number of conditional *UP-trees* recursively generated during the mining process. It is caused by the fact that *TKU* starts by a low threshold value and its strategies for raising the threshold are not very effective. Also, as *TKU* is not designed for mining over data streams, it cannot utilize the information from the past windows

Figure 7: Run Time for Phase II: (a) BMS-POS (Run time for TKU is more than 201 seconds), (b) ChainStore

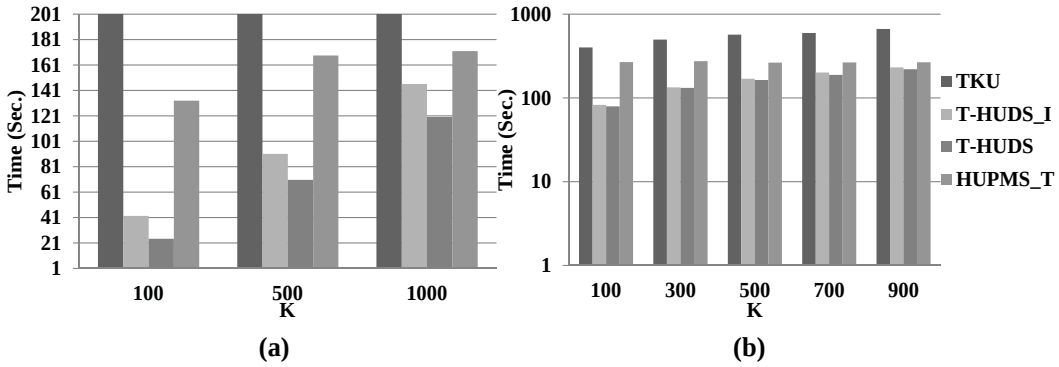


Table 4: Memory comparison (MB)

Dataset	**k**	**TKU**	**T-HUDS**	**HUPMS$_T$**
IBM	300	12.9	1.9	4.3
BMS-POS	500	37	7.5	25.2
ChainStore	300	473	75	265

to raise the threshold. In all cases, the proposed method *T-HUDS* consumes less memory than both *TKU* and *HUPMS$_T$*. Note that the node structure in *HUPMS$_T$* is also smaller than that in *T-HUDS*. But again the effective pruning strategies used in *T-HUDS* lead to generation of a smaller stack of trees in the recursive execution of the tree mining algorithm.

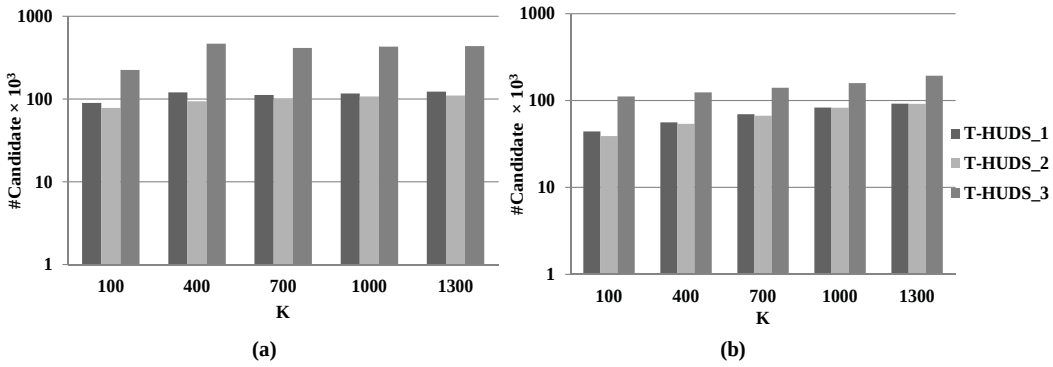
5.7. Effectiveness of the Individual Strategies

In this subsection, we investigate the impact of each of the three threshold-setting strategies used in Phase *I* of our method. Table 5 describes three different versions of the proposed method. The first method does not use

Table 5: Methods with different strategies

Method	$maxUtilList$	$MIUList$	$minTopKUtil$
$T\text{-}HUDS_1$	×	✓	✓
$T\text{-}HUDS_2$	✓	×	✓
$T\text{-}HUDS_3$	✓	✓	×

Figure 8: Number of candidates at the end of first phase for different versions of T-HUDS: (a)IBM, (b) BMS-POS datasets



maxUtillist values to set the threshold but uses $MIUList$ and the minimum top-k utility from the last window (i.e., $minTopKUtil$). *T-HUDS*$_2$ increases the threshold by means of $maxUtilList$ and $minTopKUtil$, but not by $MIUList$. *T-HUDS*$_3$ applies the first and second strategies only.

Figures 8 and 9 show the number of generated candidates and run time of these three methods on the *IBM* and *BMS-POS* datasets, respectively. In general, *T-HUDS*$_3$ (the method without the third strategy) is the worst among the three methods. It means that third strategy (i.e., using the last window's $minTopKUtil$) is the most effective strategy. *T-HUDS*$_2$ has better performance than *T-HUDS*$_1$, meaning that the first strategy (i.e., the use of $maxUtilList$) works better than the second one (i.e., using $MIUList$). Since in our implementation of TKU, $MIUList$ is used as one of the threshold-raising strategies, this results explain in part why *T-HUDS* outperforms TKU.

5.8. Effectiveness of PrefixUtil vs TWU

Below we evaluate the use of $PrefixUtil$ (in comparison to the use of TWU) for pruning the search space during the recursive tree mining process. For such a purpose, we run *T-HUDS* in the problem setting of *HUPMS*. That is, we do not use any of the threshold raising strategies in *T-HUDS* and use it as a method for finding all the high utility itemsets that satisfy an input *min_util* threshold. This is to make *T-HUDS* the same as *HUPMS* except that *T-HUDS* uses $PrefixUtil$ while *HUPMS* uses TWU to prune the search

Figure 9: Run time for different versions of T-HUDS: (a) IBM, (b) BMS-POS datasets

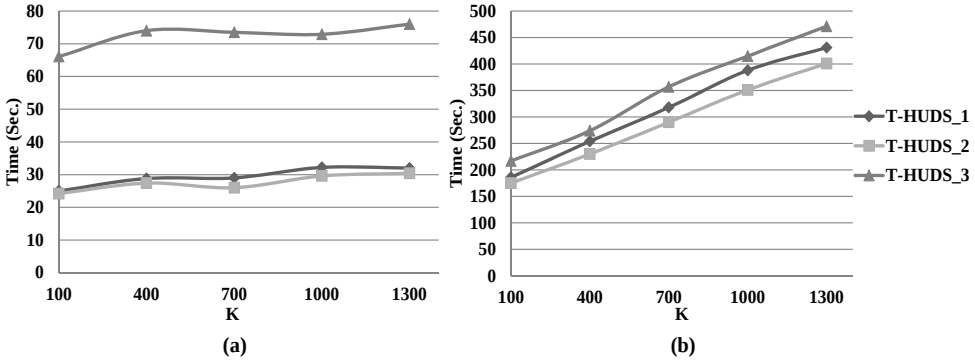


Figure 10: Impact of $PrefixUtil$ on the number of generated candidates on (a) IBM, (b) BMS-POS datasets

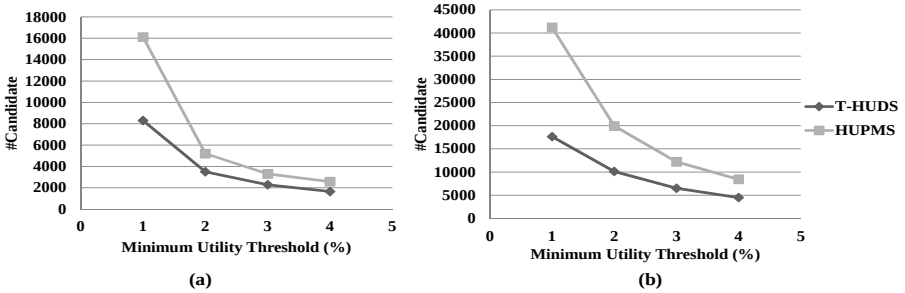


Figure 11: Impact of $PrefixUtil$ on run time on (a) IBM, (b) BMS-POS datasets

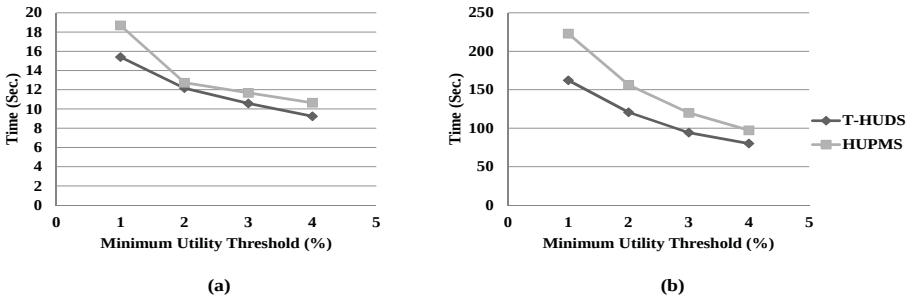
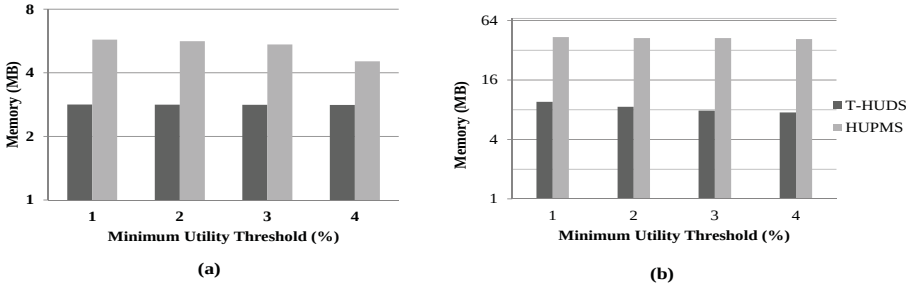


Figure 12: Impact of $PrefixUtil$ on memory comsuption on (a) IBM, (b) BMS-POS dataset



space. Hence, a comparison between these two methods will illustrate the impact of $PrefixUtil$.

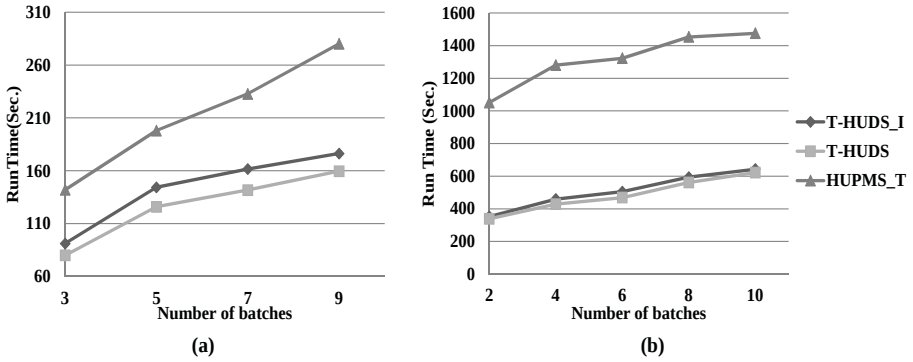
Figures 10 and 11 present the number of generated candidates in Phase one of the two methods and their total run time with different threshold values. These figures show that our algorithm significantly outperforms *HUPMS* method in terms of both the number of generated candidates and the run time. Moreover, these figures also demonstrate that the number of candidates and runtime differences increase when the minimum utility threshold decreases. As discussed earlier, the reason for $PrefixUtil$ to be more effective in pruning the search space is that it is a closer over-estimate of the true utility of an itemset than TWU.

Figure 12 shows the amount of memory consumed by two methods for mining high utility itemsets from the IBM and *BMS-POS* datasets, respectively.

5.9. T-HUDS performance with different window sizes

Because *T-HUDS* dynamically updates the tree and the set of top-k patterns once the window slides, its performance may vary depending on the window size parameter, $winSize$. In general, for a sliding window-based data stream mining algorithm, $winSize$ is an important factor on efficiency. Therefore, in order to determine the effect of changes in $winSize$ on the run time of *T-HUDS*, we analyze its performance by changing the value of this parameter. Below we present the results on the *BMS-POS* and *ChainStore* datasets, keeping the k value fixed, but changing the number of batches in the sliding window. We compare the performance of our algorithms with the $HUPMS_T$ in this experiment. Figure 13 shows the results for k = 300. The

Figure 13: Effect of the window size on the run time: (a) IBM, (b) BMS-POS datasets



y-axes in the graphs represent the overall run time (including tree construction time, update time, and mining time) for all the windows. The x-axes represent the window size in the number of batches. Each graph shows the trend in execution time with the variation of window size on each dataset. On all the $winSize$ values, the proposed method is much faster than $HUPMS_T$, and its run time increases slowly as the window size increases.

6. Related work

The MEU (Mining with Expected Utility) model [30] is the first high utility itemset mining method. MEU checks the candidate itemsets using a prediction method with a high computational cost. The $UMining$ algorithm [29] improved its performance. They defined an upper bound utility for each itemset. Using this upper bound low utility itemsets are pruned during the mining process. The *Two-Phase* method presented in [19, 20] used an over estimated utility (i.e., TWU) model for mining high utility itemsets. The main advantage of TWU is its downward closure property. In the first phase, *Two-Phase* discovers all of the high TWU itemsets ($HTWU$). Then in the second phase, it scans the database one more time to extract the true high utility itemsets from the $HTWU$ itemsets. Base on the TWU model, *CTU-Mine* [8] was proposed that is more efficient than *Two-phase* in dense databases when the minimum utility threshold is very low. This method constructs a memory-based *CUP-tree* for mining. To reduce the number of candidates in each data base scan, the isolated items discarding strategy ($IIDS$) was proposed in [17]. Applying $IIDS$, the authors proposed two

efficient algorithms *FUM* and *DCG+*. In [3], efficient tree structures were proposed to discover high utility itemset in incremental databases. This method is based on the *TWU* model as well. However, these algorithms are neither applicable to high utility itemset mining over data streams nor are able to discover top-k high utility itemsets directly.

Although several algorithms have been proposed for mining frequent patterns over data streams[5, 14, 31], these algorithms are not applicable to *HUI* mining over data streams. *THUI-Mine* [24] was the first algorithm for mining high utility itemsets from data streams. It is based on a non-stream HUI mining algorithm proposed in [19]. Later, two algorithms, called *MHUI-BIT* and *MHUI-TID*, were proposed in [15] for mining high utility itemsets from data streams. However, these methods use the Apriori-like level-wise candidate generation and thus need to scan the data in the window several times to find high utility itemsets. *GUIDE* is an algorithm proposed in [23] that mines a compact form of high utility patterns from data streams. It discovers *maximal high utility itemsets*. *HUPMS* [2] is the most recent method for HUI mining from data streams, which is based on the *TWU* model. It uses a similar data structure as we do in *T-HUDS*. However, the above mentioned methods were not designed for finding top-k high utility itemsets over data stream.

The top-k high utility itemset mining was first introduced in [4]. However, its high utility itemset definition differs from the ones used in the recently proposed methods and in ours. Recently, the *TKU* method was proposed in [27] to find top-k high utility itemsets over a static data set. The proposed approach mines top-k high utility itemset without setting the minimum utility threshold. It works based on *Up-Growth* [25]. Although it can find top-k *HUIs* effectively, it is not designed for data streams. Not only it is not able to adapt itself dynamically over different windows, but also do the proposed strategies for raising the threshold have much room to be improved so that it could generate few candidates and run faster in a data stream environment. In this paper, we designed better strategies for initializing and dynamically adjusting the minimum utility threshold over data streams.

In frequent itemset mining, several methods were proposed to find top-k frequent itemsets in static data sets [6, 7, 13, 21]. Although these algorithms are efficient, it is difficulty (if not impossible) to simply adapt them to HUI mining. There are several methods for finding top-k frequent itemsets over data streams. Golab et al. [10] proposed an algorithm, called *FREQUENT*, for the top-k frequent item discovery in sliding windows. It performs well

with bursty TPC/IP streams containing a small set of popular item types. Wong and Fu [26] present two algorithms to address the problem of top-k frequent l-itemsets ($1 \leq l \leq L$) mining over data streams. *TOPSIL-Miner* [28] is another recent algorithm for mining top-k significant itemsets over data streams, which works based on a prefix tree structure. This method is an approximation method and does not guarantee that the exact set of top-k frequent itemsets is found. A major difficulty in top-k HUI mining is that the utility of an itemset does not have the downward closure property. Thus, HUI mining has to work with estimated utilities. The strategies proposed for raising the frequency threshold in top-k frequent itemset mining do not apply to estimated utilities.

7. Conclusion

In this paper, we proposed an efficient algorithm, *T-HUDS*, for mining top-k high utility itemsets in sliding windows over streaming data. *T-HUDS* uses a novel over-estimated utility model, i.e., the *PrefixUtil* model, to effectively prune the search space for finding top-k HUIs. We prove that *PrefixUtil* satisfies a special type of the downward closure property, which allows it to be effectively used to prune the search space in a pattern growth process. We also addressed a major challenge in top-k pattern mining by devising several strategies for initializing and raising the minimum utility threshold during the mining process. A *FP-tree*-like data structure, *HUDS-tree*, and two auxiliary lists, *maxUtilList* and *MIUList*, are designed to store store the information that is needed for computing $PrefixUtil$ and for initializing and dynamically adjusting the threshold. We also designed a strategy that uses the information from the top-k patterns in the previous window to help initialize the threshold for the new window. In addition, in the second phase of top-k HUI mining, the *min_util* threshold is also raised to help fast find the top-k patterns from the candidates. We proved that using these strategies to raise the threshold and using $PrefixUtil$ to prune the search space do not miss any top-k HUIs. These strategies can not only help find top-k high utility itemsets effectively, they also reduce the run time and memory consumption of the algorithm significantly. Extensive experiments were conducted to confirm the effectiveness and the high efficiency of the algorithm in finding top-k HUIs over data streams.

While our method proves to be efficient in both run time and memory consumption, there is room for further research and improvement. Similar

to the tree structures used in *HUPMS* [2] and *TKU* [27], the *HUDS-tree* is a lossy compression of the transactions in the sliding window. The consequence of this is that a second scan of data in the sliding window is needed in the second phase of the method to obtain the exact utilities of the potential top-k HUIs. Although a sliding window is generally small enough to fit into the main memory, reducing the number of data scans can further improve the run time performance of HUI mining. We will look into two directions: one is to design a lossless compression data structure to store the information needed to compute the exact utility, and the other is to design an approximation method that returns an approximate list of top-k patterns from a lossy compression of the data.

References

[1] R. Agrawal, R. Srikant, Fast algorithms for mining association rules, in: Proceedings of the 20th International Conference on Very Large Data Bases, 1994, pp. 487–499.

[2] C.F. Ahmed, S.K. Tanbeer, B.S. Jeong, Interactive mining of high utility patterns over data streams, Expert Systems with Applications 39 (2012) 11979–11991.

[3] C.F. Ahmed, S.K. Tanbeer, B.S. Jeong, Y.K. Lee, Efficient tree structures for high-utility pattern mining in incremental databases, IEEE Transactions on Knowledge and Data Engineering 21 (2009) 1708–1721.

[4] R. Chan, Q. Yang, Y. Shen, Mining high-utility itemsets, in: Proc. of Third IEEE Int'l Conf. on Data Mining, 2003, pp. 19–26.

[5] J. Cheng, Y. Ke, W. Ng, A survey on algorithms for mining frequent itemsets over data streams, Knowledge and Information Systems 16 (2008) 1–27.

[6] Y.L. Cheung, A.W. Fu, Mining frequent itemsets without support threshold: with and without item constraints, IEEE Transactions on Knowledge and Data Engineering 16 (2004) 1052–1069.

[7] K. Chuang, J. Huang, M. Chen, Mining top-k frequent patterns in the presence of the memory constraint, The VLDB Journal 17 (2008) 1321–1344.

[8] A. Erwin, R.P. Gopalan, N.R. Achuthan, A bottom-up projection based algorithm for mining high utility itemsets, in: Proceedings of 2nd International Workshop on Integration Artificial Intelligence and Data Mining, 2007, pp. 3–11.

[9] B. Goethals, M.J. Zaki., Frequent itemset mining dataset repository, http://fimi.cs.helsinki.fi/data/, 2004.

[10] L. Golab, D. Dehaan, E. Demaine, Identifying frequent items in sliding windows over on-line packet streams, in: Proceedings of ACM SIGCOMM internet measurement conference, 2003, pp. 173–178.

[11] J. Han, H. Cheng, D. Xin, X. Yan, Frequent pattern mining: current status and future directions, Data Mining and Knowledge Discovery 15 (2007) 55–86.

[12] J. Han, J. Pei, Y. Yin, Mining frequent patterns without candidate generation, SIGMOD Rec. 29 (2000) 1–12.

[13] Y. Hirate, E. Iwahashi, H. Yamana, Tf2p-growth: An efficient algorithm for mining frequent patterns without any thresholds, in: Proc. of IEEE ICDM'04 Workshop on Alternative Techniques for Data Mining and Knowledge Discoverey, 2004.

[14] K.S.C. Leung, F. Jiang, Frequent itemset mining of uncertain data streams using the damped window model, in: Proceedings of the 2011 ACM Symposium on Applied Computing, 2011, pp. 950–955.

[15] H.F. Li, H.Y. Huang, Y.C. Chen, Y.J. Liu, S.Y. Lee, Fast and memory efficient mining of high utility itemsets in data streams, in: Proc. of the 8th IEEE Int'l Conf. on Data Mining, 2008, pp. 881–886.

[16] H.F. Li, H.Y. Huang, S.Y. Lee, Fast and memory efficient mining of high-utility itemsets from data streams: with and without negative item profits, Knowledge and Information Systems 28 (2011) 495–522.

[17] Y.C. Li, J.S. Yeh, C.C. Chang, Isolated items discarding strategy for discovering high utility itemsets, Data and Knowledge Engineering 64 (2008) 198–217.

[18] M. Liu, J. Qu, Mining high utility itemsets without candidate generation, in: Proceedings of the 21st ACM international conference on Information and knowledge management, 2012, pp. 55–64.

[19] Y. Liu, W. k. Liao, A. Choudhary, A fast high utility itemsets mining algorithm, in: Proceedings of the 1st international workshop on Utility-based data mining, 2005, pp. 90–99.

[20] Y. Liu, W. Liao, A. Choudhary, A two-phase algorithm for fast discovery of high utility of itemsets, in: Proceedings of the 9th Pacific-Asia Conference on Knowledge Discovery and Data Mining, 2005, pp. 689–695.

[21] S. Ngan, T. Lam, R.C. Wong, A.W. Fu, Mining n-most interesting itemsets without support threshold by the cofi-tree, Int. J. Business Intelligence and Data Mining 1 (2005) 88–106.

[22] J. Pisharath, Y. Liu, B. Ozisikyilmaz, R. Narayanan, W.K. Liao, A. Choudhary, G. Memik, Nu-minebench version 2.0 dataset and technical report, http://cucis.ece.northwestern.edu/projects/dms/minebench.html, 2012.

[23] B.E. Shie, V.S. Tseng, P.S. Yu, Online mining of temporal maximal utility itemsets from data streams, in: Proceedings of the 2010 ACM Symposium on Applied Computing, 2010, pp. 1622–1626.

[24] V.S. Tseng, C.J. Chu, T. Liang, Efficient mining of temporal high-utility itemsets from data streams, in: ACM KDD Utility Based Data Mining, 2006, pp. 18–27.

[25] V.S. Tseng, C.W. Wu, B.E. Shie, P.S. Yu, Up-growth: an efficient algorithm for high utility itemset mining, in: Proc. of Int'l Conf. on ACM SIGKDD, 2010, pp. 253–262.

[26] R.C.W. Wong, A.W.C. Fu, Mining top-k frequent itemsets from data streams, Data Mining and Knowledge Discovery 13 (2006) 193–217.

[27] C.W. Wu, B.E. Shie, V.S. Tseng, P.S. Yu, Mining top-k high utility itemsets, in: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, 2012, pp. 78–86.

[28] B. Yang, H. Huang, Topsil-miner:an efficient algorithm for mining top-k significant itemsets over data streams, Data Mining and Knowledge Discovery 23 (2010) 225–242.

[29] H. Yao, H.J. Hamilton, Mining itemset utilities from transaction databases, Data and Knowledge Engineering 59 (2006) 603–626.

[30] H. Yao, H.J. Hamilton, C.J. Butz, A foundational approach to mining itemset utilities from database, in: Proceeding of the 4th SIAM International Conference on Data Mining, 2004, pp. 482–491.

[31] S.J. Yen, Y.S. Lee, C.W. Wu, C.L. Lin, An efficient algorithm for maintaining frequent closed itemsets over data stream, in: Proceedings of IEA/AIE, 2009, pp. 767–776.