This is the peer reviewed version of the following article: Feng Ml et al. 2010, 'Pattern Space Maintenance For Data Updates And Interactive Mining', Wiley-blackwell, vol. 26, no. 3, pp. 282-317. which has been published in final form at http://dx.doi.org/10.1111/j.1467-8640.2010.00360.x This article may be used for non-commercial purposes in accordance With Wiley Terms and Conditions for self-archiving'

Pattern Space Maintenance for Data Updates & Interactive Mining *

Mengling Feng †[1] Guozhu Dong,[2] Jinyan Li,[1]
Yap-Peng Tan,[1] Limsoon Wong [3]
[1] Nanyang Technological University,[2]Wright State University
[2]National University of Singapore

Abstract

This paper addresses the incremental and decremental maintenance of the frequent pattern space. We conduct an in-depth investigation on how the frequent pattern space evolves under both incremental and decremental updates. Based on the evolution analysis, a new data structure, *Generator-Enumeration Tree* (*GE-tree*), is developed to facilitate the maintenance of the frequent pattern space. With the concept of *GE-tree*, we propose two novel algorithms, *Pattern Space Maintainer+* (PSM+) and *Pattern Space Maintainer-* (PSM-), for the incremental and decremental maintenance of frequent patterns. Experimental results demonstrate that the proposed algorithms, on average, outperforms the representative state-of-the-art methods by an order of magnitude.

Key words: Data Mining, Frequent Pattern, Incremental Maintenance, Data Update & Interactive Mining.

*This manuscript is to be submitted to the Special Issue of Computational Intelligence on "Advanced Data Mining and Applications". The conference version of this paper was published in ADMA'2008. Editors: Charles Ling and Qiang Yang (Guest Editors).

†Corresponding author. Email: mornin@gmail.com

1 Introduction

Updates are a fundamental aspect of data management. Updates allow obsolete and incorrect records to be removed and new records to be included. When a database is updated frequently, repeating the pattern discovery process from scratch during each update causes significant computational and I/O overheads. Therefore, it is important to analyze how the discovered pattern may change in response to updates, and to formulate more effective algorithms to maintain the discovered pattern on the updated database.

Pattern maintenance is also useful for interactive mining applications. For example, pattern maintenance can be used to interactively analyze the evolution trend of a time series data. This type of trend analysis usually focuses on a certain period of time, and patterns found before the targeted period are first extracted as a reference. Then records within the targeted period are inserted one by one in time sequence. The patterns before and after the insertion are then compared to find out whether any new patterns (trends) have emerged and how the existing patterns (trends) have changed. This interactive study is a useful tool to detect significant events, like the emergence of new trend, changes of the existing trends, vanishing trends, etc. More importantly, through the study, we can also identify the time when the significant events happened, which allows further investigation on the causes of the events. This type of "before vs. after" analysis requires intensive pattern discovery and comparison computation. Solving the problem using the conventional pattern discovery methods involves large amount of redundancies, and pattern maintenance can be used to effectively avoid these redundancies.

This paper addresses the maintenance of the frequent patterns space. Frequent patterns (Agrawal and Imielinski, 1993) are a very important type of patterns in data mining. Frequent patterns play an essential role in various

knowledge discovery tasks, such as the discovery of association rules, correlations, causality, sequential patterns, emerging patterns, etc. The frequent patterns space, consisting all the frequent patterns, is usually very large. Thus the maintenance of the frequent pattern space is computational challenging.

In this paper, we focus on two major types of updates in data management and interactive mining. The first type, where new transactions are inserted into the original dataset, is called an *incremental update*. The associated maintenance process is called *incremental maintenance*. The second type, where some transactions are removed from the original dataset, is called a *decremental update*. The associated maintenance process is called *decremental maintenance*.

1.1 Related Work

In the literature, the frequent pattern maintenance algorithms can be classified into four main categories: the 1) *Apriori-based* algorithms, 2) *Partition-based* algorithms, 3) *Prefix-tree-based* algorithms and 4) *Concise-representation-based* algorithms.

FUP (Cheung *et al.*, 1996) is the first *Apriori*-based maintenance algorithm. FUP focuses on the incremental maintenance of frequent patterns. Inspired by Apriori (Agrawal and Imielinski, 1993), FUP updates the space of frequent patterns iteratively based on the candidate-generation-verification framework. The key technique of FUP is to makes use of support information in previously discovered frequent patterns to minimize the number of candidate patterns. Since the performance of candidate-generation-verification based algorithms heavily depends on the size of the candidate set, FUP outperforms Apriori. FUP is then generalized as FUP2H (Cheung *et al.*, 1997) to handle both incremental and decremental maintenance. Similarly, the partition-based algorithm SWF

(Lee *et al.*, 2005) also employs the candidate-generation-verification framework. However, SWF applies different techniques to reduce the size of candidate set. SWF slices a dataset into several partitions and employs a filtering threshold in each partition to filter out unnecessary candidate patterns. Even with all the candidate reduction techniques, the candidate-generation-verification framework still leads to the enumeration of large number of unnecessary candidates. This greatly limits the performance of both *Apriori*-based and partition-based algorithms.

To address this shortcoming of the candidate-generation-verification framework, prefix-tree-based algorithms, such as CanTree (Leung *et al.*, 2007), that involve no candidate generation are proposed. CanTree evolves from FP-growth (Han *et al.*, 2000) — the state-of-the-art prefix-tree-based frequent pattern discovery algorithm. CanTree arranges items according to some fixed canonical order that will not be affected by data updates. This allows new transactions to be efficiently inserted into the existing prefix-tree without node swapping/merging. However, prefix-tree based algorithms still suffer from the undesirably large size of the frequent pattern space.

To break this bottleneck, concise representations of the frequent pattern space are proposed. The commonly used representations include "maximal patterns" (Bayardo, 1998), "closed patterns" and "generators" (Pasquier *et al.*, 1999). Algorithms have also been proposed to maintain the concise representations. Moment (Chi *et al.*, 2006) is one example. Moment dynamically maintains the frequent closed patterns. Moment focuses on a special update scenario where each time only one new transaction is inserted and one obsolete transaction is removed, and thus it is proposed based on the hypothesis that there are only *small changes* to the frequent closed patterns given a small amount of updates. Due to this unfavorable constraint, the performance of Moment degrades dra-

matically when the number of updates gets large. ZIGZAG (Veloso *et al.*, 2002), on the other hand, maintains the maximal patterns. Extended from the maximal pattern discovery algorithm GENMAX (Gouda and Zaki, 2001), ZIGZAG updates the maximal patterns by a backtracking search, which is guided by the outcomes of the previous maintenance iteration. However, the maximal patterns are a lossy representation of the frequent pattern space, which do not provide support information of frequent patterns.

We observe that most of the prior works in frequent pattern maintenance, e.g. FUP, CanTree and ZIGZAG, are proposed as an extension of frequent pattern discovery algorithms. Unlike these prior works, we propose our maintenance algorithms based on an in-depth analysis on the evolution of the pattern space under data updates. The evolution of the pattern space is analyzed using the concept of equivalence classes. Different from the maximal pattern in ZIGZAG, the equivalence class is a lossless [1] concise representation of the frequent pattern space. Also, unlike Moment, which bears some unfavorable assumptions, our maintenance algorithms aim to handle batch updates.

1.2 Our Contribution

Our contributions in this paper are as follows. (1) We analyze how the space of frequent patterns evolves under both incremental and decremental updates. Based on this space evolution analysis, we summarize the major computation tasks involved in the frequent pattern maintenance. (2) To effectively address the maintenance computational tasks, we develop a data structure, *Generator-Enumeration Tree* (*GE-tree*). Inspired by the idea of *Set-Enumeration Tree* (*SE-tree*) (Rymon, 1992), *GE-tree* efficiently facilitates the frequent pattern maintenance. (3) We propose two novel maintenance algorithms, *Pattern Space*

[1] We say a representation is lossless if it is sufficient to derive and determine the support of all frequent patterns without accessing the datasets.

Maintainer+ (PSM+) and *Pattern Space Maintainer-* (PSM-). With *GE-tree*, PSM+ and PSM- effectively maintain the frequent pattern space under incremental and decremental updates. (4) We also demonstrate that PSM+ and PSM- can be easily integrated to form *Pattern Space Maintainer* (PSM), and PSM can be extended to update the frequent pattern space for support threshold adjustment. (5) We have conducted extensive experiments to evaluate the effectiveness of our proposed algorithms. Experimental results show that the proposed algorithms, on average, outperform the state-of-the-art approaches by more than an order of magnitude.

The rest of the paper is organized as follows. In Section 2, we recap the basic definitions in frequent pattern maintenance. In Section 3, we investigate how the space of frequent pattern can be structurally decomposed into and represented by equivalence classes. In Section 4 and 5, we discuss the proposed incremental and decremental maintenance algorithms. The generalized and extension of the proposed algorithms are discussed in Section 6, and the experimental results are presented in Section 7. We conclude the paper in Section 8.

2 Problem Definition

Let $\mathcal{I} = \{i_1, i_2, ..., i_m\}$ be a set of distinct literals called "items", and also let $\mathcal{D} = \{t_1, t_2, ..., t_n\}$ be a transactional "dataset", where t_i ($i \in [1, n]$) is a "transaction" that contains a non-empty set of items. Each subset of $\mathcal{I}$ is called a "pattern" or an "itemset". The "support" of a pattern P in a dataset $\mathcal{D}$ is defined as $sup(P, \mathcal{D}) = |\{t | t \in \mathcal{D} \wedge P \subseteq t\}|$. A pre-specified support threshold is necessary to define frequent patterns. The support threshold can be defined in terms of percentage and absolute count. For a dataset $\mathcal{D}$, the "percentage support threshold", $ms_{\%}$, and the "absolute support threshold", ms_a, can be interchanged via equation $ms_a = \lceil ms_{\%} \times |\mathcal{D}| \rceil$. For this paper, we assume the

percentage support threshold is used unless otherwise specified. Given $ms_{\%}$ or ms_a, a pattern P is said to be *frequent* in a dataset $\mathcal{D}$ iff $sup(P, \mathcal{D}) \geq ms_a = \lceil ms_{\%} \times |\mathcal{D}| \rceil$. The collection of all frequent patterns in $\mathcal{D}$ is called the "space of frequent patterns" or the "frequent pattern space" and is denoted as $\mathcal{F}(\mathcal{D}, ms_{\%})$ or $\mathcal{F}(\mathcal{D}, ms_a)$.

For incremental maintenance, we use the following notations: $\mathcal{D}_{org}$ is the original dataset, $\mathcal{D}_{inc}$ is the set of new transactions to be added to $\mathcal{D}_{org}$, and $\mathcal{D}_{upd+} = \mathcal{D}_{org} \cup \mathcal{D}_{inc}$ is the updated dataset. We assume without loss of generality that $\mathcal{D}_{org} \cap \mathcal{D}_{inc} = \emptyset$. This leads to the conclusion that $|\mathcal{D}_{upd+}| = |\mathcal{D}_{org}| + |\mathcal{D}_{inc}|$. Given $ms_{\%}$, the task of incremental maintenance is to obtain the updated frequent pattern space $\mathcal{F}(\mathcal{D}_{upd+}, ms_{\%})$ by updating the original pattern space $\mathcal{F}(\mathcal{D}_{org}, ms_{\%})$.

On the other hand, we use the following notations for decremental maintenance: $\mathcal{D}_{dec}$ is the set of old transactions to be removed, and $\mathcal{D}_{upd-} = \mathcal{D}_{org} - \mathcal{D}_{dec}$ is the updated dataset. We assume without loss of generality that $\mathcal{D}_{dec} \subseteq \mathcal{D}_{org}$. Thus $|\mathcal{D}_{upd-}| = |\mathcal{D}_{org}| - |\mathcal{D}_{dec}|$. Given $ms_{\%}$, the task of decremental maintenance is to obtain the updated frequent pattern space $\mathcal{F}(\mathcal{D}_{upd-}, ms_{\%})$ by updating the original pattern space $\mathcal{F}(\mathcal{D}_{org}, ms_{\%})$.

3 Structural Decomposition of Pattern Space

Understanding how the frequent pattern space evolves when data is updated is essential for effective maintenance of the space. However, due to the vast size of the frequent pattern space, direct analysis on the pattern space is extremely difficult. To solve this problem, we propose to structurally decompose the frequent pattern space into sub-spaces.

We observe that the frequent pattern space is a convex space.

Definition 3.1 (Convex Space) *A space $\mathcal{S}$ is convex if, for all $X, Y \in \mathcal{S}$ such*

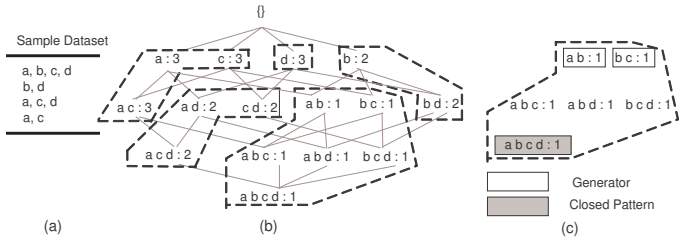


Figure 1: Demonstration of the structural decomposition of the frequent pattern space. (a)The sample dataset; (b) decomposition of the frequent pattern space of the sample dataset into 5 equivalence classes; (c) the "border" of an equivalence class.

that $X \subseteq Y$, it is the case that $Z \in S$ whenever $X \subseteq Z \subseteq Y$.

For a convex space $\mathcal{S}$, we define the collection of all "most general" patterns in $\mathcal{S}$ as a "bound" of $\mathcal{S}$. A pattern X is most general in $\mathcal{S}$ if there is no proper subset of X in $\mathcal{S}$. Similarly, we define the collection of all "most specific" patterns as another bound of $\mathcal{S}$. A pattern X is most specific in $\mathcal{S}$ if there is no proper superset of X in $\mathcal{S}$. We call the former bound the "left bound" of $\mathcal{S}$, denoted $\mathcal{L}$; and the latter bound the "right bound" of $\mathcal{S}$, denoted $\mathcal{R}$. We call the pair of left and right bound the "border" of $\mathcal{S}$, which is denoted by $\langle \mathcal{L}, \mathcal{R} \rangle$. It is easy to show that a convex space can be concisely represented by its borders without loss of information.

Fact 3.2 (Cf. Li *et al.* (2005)) *$\mathcal{F}(ms_{\%}, \mathcal{D})$ is convex. Furthermore, it can be structurally decomposed into convex sub-spaces — equivalence classes.*

We further found that, due to its convexity, the frequent pattern space can be structurally decomposed into sub-spaces, which is mush smaller in terms of size. The sub-space is called the equivalence class, and it is formally defined as follows.

Definition 3.3 (Equivalence Class) *Let the "filter", $f(P, \mathcal{D})$, of a pattern P in a dataset $\mathcal{D}$ be defined as $f(P, \mathcal{D}) = \{T \in \mathcal{D} \mid P \subseteq T\}$. Then the "equivalence*

class" $[P]_{\mathcal{D}}$ *of* P *in a dataset* $\mathcal{D}$ *is the collection of patterns defined as* $[P]_{\mathcal{D}} = \{Q \mid f(P,\mathcal{D}) = f(Q,\mathcal{D})$*,* Q *is a pattern in* $\mathcal{D}\}$.

In other words, two patterns are "equivalent" in the context of a dataset $\mathcal{D}$ iff they are included in exactly the same transactions in $\mathcal{D}$. Thus the patterns in a given equivalence class have the same support. So we extend the notations and write $sup(P,\mathcal{D})$ to denote the support of an equivalence class $[P]_{\mathcal{D}}$ and $P \in \mathcal{F}(ms,\mathcal{D})$ to mean the equivalence class is frequent. Furthermore, equivalence classes are also convex and thus they can be compactly represented by their borders without loss of information (Li *et al.*, 2005). The right bound of an equivalence class is actually a closed pattern, and the left bound is a group of generators (key patterns).

Definition 3.4 (Generator & Closed Pattern (Pasquier *et al.*, 1999)) *A pattern* P *is a "key pattern" or a "generator" in a dataset* $\mathcal{D}$ *iff for every* $P' \subset P$*, it is the case that* $sup(P',\mathcal{D}) > sup(P,\mathcal{D})$*. In contrast, a pattern* P *is a "closed pattern" in a dataset* $\mathcal{D}$ *iff for every* $P' \supset P$*, it is the case that* $sup(P',\mathcal{D}) < sup(P,\mathcal{D})$.

Based on the definition of the border of a convex space, we can define generators and closed patterns in an alternative way.

Fact 3.5 *A pattern* P *is a key pattern or a generator in a dataset* $\mathcal{D}$ *iff* P *is a most general pattern in* $[P]_{\mathcal{D}}$*. A pattern* P *is a closed pattern in a dataset* $\mathcal{D}$ *iff* P *is the most specific pattern in* $[P]_{\mathcal{D}}$.

Therefore, the closed pattern and generators form the border of the corresponding equivalence class, and they, furthermore, uniquely define the corresponding equivalence class. This implies that, to mine or maintain generators and closed patterns, it is sufficient to mine or maintain the borders of equivalence classes, and vice versa. Figure 1 graphically demonstrates how the pattern space

can be structurally decomposed into equivalence classes and how an equivalence class can be concisely represented by its border.

In addition, we observe that generators follow the "a priori" (or anti-monotone) property.

Fact 3.6 (Cf. Li *et al.* (2005)) *Let P be a pattern in $\mathcal{D}$. If P is frequent, then every subset of P is also frequent. If P is a generator, then every subset of P is also a generator in $\mathcal{D}$. Thus, if P is a frequent generator, then every subset of P is also a frequent generator in $\mathcal{D}$.*

The equivalence class is an effective concise representation for pattern spaces. In the literature, the equivalence class has been used to summarize cells in data cubes (Li *et al.*, 2004). Here we use equivalence classes to concisely represent the space of frequent patterns. Structurally decomposing the pattern space into equivalence classes allows us to investigate the evolution of the pattern space via studying the evolution of equivalence classes, which is much smaller and easier to study. Moreover, the structural decomposition simplifies the maintenance problem from updating the entire space to the update of equivalence classes, and it also allows us to maintain the pattern space in a divide-and-conquer manner.

4 Incremental Maintenance of Pattern Space

This section discusses the incremental maintenance of the frequent pattern space. In the incremental update, a set of new transactions $\mathcal{D}_{inc}$ are inserted into the original dataset $\mathcal{D}_{org}$, and thus the updated dataset $\mathcal{D}_{upd+} = \mathcal{D}_{org} \cup \mathcal{D}_{inc}$. Given a support threshold $ms_{\%}$, the task of incremental maintenance is to obtain the updated pattern space by maintaining the original pattern space.

To develop effective incremental maintenance algorithm, we start off with a

study on the evolution of the frequent pattern space under incremental updates using the concept of equivalence class. Through the space evolution study, we summarize the major computational tasks in the incremental maintenance. To complete the computational tasks efficiently, we develop a new data structure, *Generator-Enumeration Tree* (*GE-tree*). Based on the *GE-tree*, a novel incremental maintenance algorithm, named *Pattern Space Maintainer+* (PSM+), is proposed.

4.1 Evolution of Pattern Space

We first investigate how the existing (frequent) equivalence classes evolve when new transactions are added. We observe that, after an incremental update, the support of an equivalence class can only increase and the size of an equivalence class can only shrink.

Proposition 4.1 *Let P be a pattern in $\mathcal{D}_{org}$. Then $[P]_{\mathcal{D}_{upd+}} \subseteq [P]_{\mathcal{D}_{org}}$ and $sup(P, \mathcal{D}_{upd+}) \geq sup(P, \mathcal{D}_{org})$.*

Proof: *Suppose $Q \in [P]_{\mathcal{D}_{upd+}}$. Then $f(Q, \mathcal{D}_{upd+}) = f(Q, \mathcal{D}_{org}) \cup f(Q, \mathcal{D}_{inc}) = f(P, \mathcal{D}_{upd+}) = f(P, \mathcal{D}_{org}) \cup f(P, \mathcal{D}_{inc})$. Since $\mathcal{D}_{inc} \cup \mathcal{D}_{org} = \emptyset$, we have $f(Q, \mathcal{D}_{org}) = f(P, \mathcal{D}_{org})$. Then $Q \in [P]_{\mathcal{D}_{org}}$ for every $Q \in [P]_{\mathcal{D}_{upd+}}$. Thus we can conclude $[P]_{\mathcal{D}_{upd+}} \subseteq [P]_{\mathcal{D}_{org}}$. Also, $sup(P, \mathcal{D}_{upd+}) = sup(P, \mathcal{D}_{org}) + sup(P, \mathcal{D}_{inc}) \geq sup(P, \mathcal{D}_{org})$.* □

In particular, we discover that, under an incremental update, the existing equivalence classes evolve in three different ways. The first way is to remain unchanged without any change in support. The second way is to remain unchanged but with an increased support. The third way is to split into two or more classes. In this case, the size of equivalence classes will shrink as described in Proposition 4.1. On the other hand, an incremental update may induce new [2]

[2]We can an equivalence class is "new" iff the patterns in the class are not in the original pattern space but in the updated pattern space.

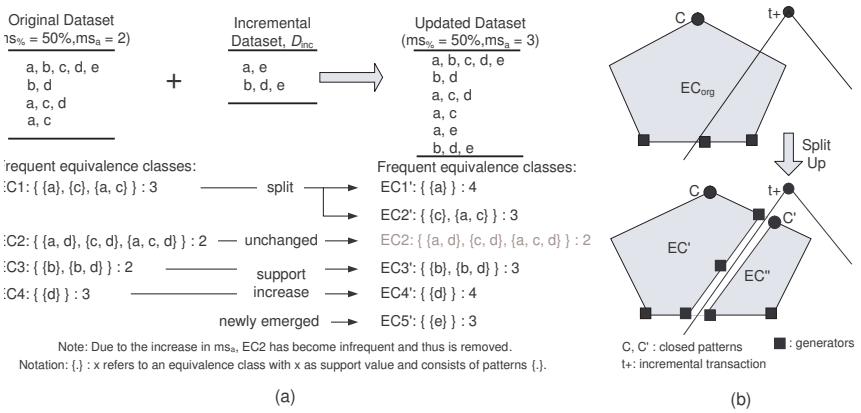


Figure 2: (a)The evolution of the frequent pattern space under the incremental update; (b) the splitting up of an equivalence class EC_{org} after $t+$ is inserted.

(frequent) equivalence classes to emerge. Figure 2 (a) illustrates how the existing equivalence classes may evolve in three different ways and how new equivalence classes may emerge.

To have an in-depth understanding on how the pattern space evolve under the incremental update, we now investigate the exact conditions for the three ways that existing equivalence classes may evolve and also the conditions for new equivalence classes to emerge. We denote the closed pattern of an equivalence class $[p]_{\mathcal{D}}$ as $Clo([p]_{\mathcal{D}})$ and the generators or key patterns of $[p]_{\mathcal{D}}$ as $Keys([p]_{\mathcal{D}})$. We assume the incremental dataset $\mathcal{D}_{inc}$ contains only one transaction t_+ for ease of discussion.

Theorem 4.2 *Let $\mathcal{D}_{org}$ be the original dataset, $\mathcal{D}_{inc}$ be the incremental dataset, $\mathcal{D}_{upd+} = \mathcal{D}_{org} \cup \mathcal{D}_{inc}$ and $ms_{\%}$ be the support threshold. Suppose $\mathcal{D}_{inc}$ consists of only one transaction t_+. For every frequent equivalence class $[P]_{\mathcal{D}_{upd+}}$ in $\mathcal{F}(ms_{\%}, \mathcal{D}_{upd+})$, exactly one of the 5 scenarios below holds:*

1. *$P \in \mathcal{F}(ms_{\%}, \mathcal{D}_{org})$, $P \not\subseteq t_+$ and $Q \not\subseteq t_+$ for all $Q \in [P]_{\mathcal{D}_{org}}$, corresponding to the scenario where the equivalence class remains totally unchanged. In*

this case, $[P]_{\mathcal{D}_{upd+}} = [P]_{\mathcal{D}_{org}}$ *and* $sup(P, \mathcal{D}_{upd+}) = sup(P, \mathcal{D}_{org})$.

2. $P \in \mathcal{F}(ms_{\%}, \mathcal{D}_{org})$, $P \subseteq t_+$ *and* $Q \subseteq t_+$ *for all* $Q \in [P]_{\mathcal{D}_{org}}$, *corresponding to the scenario where the equivalence class has remained unchanged but with increased support. In this case,* $[P]_{\mathcal{D}_{upd+}} = [P]_{\mathcal{D}_{org}}$ *and* $sup(P, \mathcal{D}_{upd+}) = sup(P, \mathcal{D}_{org}) + sup(P, \mathcal{D}_{upd+}) = sup(P, \mathcal{D}_{org}) + 1$.

3. $P \in \mathcal{F}(ms_{\%}, \mathcal{D}_{org})$, $P \subseteq t_+$ *and* $Q \not\subseteq t_+$ *for some* $Q \in [P]_{\mathcal{D}_{org}}$, *corresponding to the scenario where the equivalence class splits. In this case,* $[P]_{\mathcal{D}_{org}}$ *splits into two new equivalence classes, and* $[P]_{\mathcal{D}_{upd+}}$ *is one of them.* $[P]_{\mathcal{D}_{upd+}} = \{Q | Q \in [P]_{\mathcal{D}_{org}} \wedge Q \subseteq t_+\}$, $Clo([P]_{\mathcal{D}_{upd+}}) = Clo([P]_{\mathcal{D}_{org}}) \cap t_+$ *and* $Keys([P]_{\mathcal{D}_{upd+}}) = \{K | K \in Keys([P]_{\mathcal{D}_{org}}) \wedge K \subseteq t_+\}$.

4. $P \in \mathcal{F}(ms_{\%}, \mathcal{D}_{org})$, $P \not\subseteq t_+$ *and* $Q \subseteq t_+$ *for some* $Q \in [P]_{\mathcal{D}_{org}}$, *also corresponding to the scenario where the equivalence class splits. This scenario is complement to Scenario 3.* $[P]_{\mathcal{D}_{org}}$ *splits into two new equivalence classes,* $[P]_{\mathcal{D}_{upd+}}$ *is one of them, and the other one has been described in Scenario 3. In this case,* $[P]_{\mathcal{D}_{upd+}} = \{Q | Q \in [P]_{\mathcal{D}_{org}} \wedge Q \not\subseteq t_+\}$, $Clo([P]_{\mathcal{D}_{upd+}}) = Clo([P]_{\mathcal{D}_{org}})$ *and* $Keys([P]_{\mathcal{D}_{upd+}}) = min\{\{K | K \in Keys([P]_{\mathcal{D}_{org}}) \wedge K \not\subseteq t_+\} \cup \{K' \cup \{x_i\}, i = 1, 2, \cdots | K' \in Keys([P]_{\mathcal{D}_{org}}) \wedge K' \subseteq t_+, x_i \in Clo([P]_{\mathcal{D}_{org}}) \wedge x_i \notin t_+\}\}$.

5. $P \notin \mathcal{F}(ms_{\%}, \mathcal{D}_{org})$, $P \subseteq t_+$ *and* $Sup(P, \mathcal{D}_{upd+}) \geq \lceil ms_{\%} \times |\mathcal{D}_{upd+}| \rceil$, *corresponding to the scenario where a new frequent equivalence class has emerged. In this case,* $[P]_{\mathcal{D}_{upd+}} = \{Q | Q \in [P]_{\mathcal{D}_{org}} \wedge Q \subseteq t_+\}$ *and* $sup(P, \mathcal{D}_{upd+}) = sup(P, \mathcal{D}_{org}) + sup(P, \mathcal{D}_{upd+}) = sup(P, \mathcal{D}_{org}) + 1$.

Proof: *Scenario 1 and 5 are obvious.*

To prove Scenario 2, suppose (i) $P \in \mathcal{F}(ms_{\%}, \mathcal{D}_{org})$, *(ii)* $P \subseteq t_+$ *and (iii)* $Q \subseteq t_+$ *for all* $Q \in [P]_{\mathcal{D}_{org}}$. *Point (ii) implies that* $f(P, \mathcal{D}_{upd+}) = f(P, \mathcal{D}_{org}) \cup$

$\{t_+\}$, *and point (iii) implies that, for all* $Q \in [P]_{\mathcal{D}_{org}}$, $f(Q, \mathcal{D}_{upd+}) = f(Q, \mathcal{D}_{org}) \cup \{t_+\}$. *Since, according to the definition of equivalence class (Definition 3.3),* $f(P, \mathcal{D}_{org}) = f(Q, \mathcal{D}_{org})$. *Thus* $f(P, \mathcal{D}_{upd+}) = f(P, \mathcal{D}_{org}) \cup \{t_+\} = f(Q, \mathcal{D}_{org}) \cup \{t_+\} = f(Q, \mathcal{D}_{upd+})$. *This means that, for all* $Q \in [P]_{\mathcal{D}_{org}}$, $Q \in [P]_{upd+}$. *Therefore, the equivalence* $[P]_{\mathcal{D}_{org}}$ *remains the same after the update, but* $sup(P, \mathcal{D}_{upd+}) = |f(P, \mathcal{D}_{upd+})| = sup(P, \mathcal{D}_{org}) + 1$.

To prove Scenario 3, suppose (i) $P \in \mathcal{F}(ms_{\%})$, *(ii)* $P \subset t_+$, *and (iii)*$Q \not\subseteq t_+$ *for some* $Q \in [P]_{\mathcal{D}_{org}}$. *Point (ii) implies that* $f(P, \mathcal{D}_{upd+}) = f(P, \mathcal{D}_{org}) \cup \{t_+\}$. *Also for patterns* Q *that satisfy point (iii),* $f(Q, \mathcal{D}_{upd+}) = f(Q, \mathcal{D}_{org}) \neq f(P, \mathcal{D}_{upd+})$. *This means* $Q \notin [P]_{\mathcal{D}_{upd+}}$. *According to Definition 3.3,* $[P]_{\mathcal{D}_{upd+}} = \{P' | f(P, \mathcal{D}_{upd+}) = f(P', \mathcal{D}_{upd+})\} = \{P' | P' \in [P]_{\mathcal{D}_{org}} \wedge P' \subseteq t_+\}$, *and* $[Q]_{\mathcal{D}_{upd+}} = \{Q' | Q' \in [P]_{\mathcal{D}_{org}} \wedge Q' \not\subseteq t_+\}$. *Since* $[P]_{\mathcal{D}_{org}} = [P]_{\mathcal{D}_{upd+}} \cup [Q]_{\mathcal{D}_{upd+}}$ *and* $[P]_{\mathcal{D}_{upd+}} \cap [Q]_{\mathcal{D}_{upd+}} = \emptyset$, *we say that, in this case, the equivalence class* $[P]_{\mathcal{D}_{org}}$ *splits into two.*

Next, we prove $Clo([P]_{\mathcal{D}_{upd+}}) = Clo([P]_{\mathcal{D}_{org}}) \cap t_+$. *Let* $C = Clo([P]_{\mathcal{D}_{org}}) \cap t_+$. *It is obvious that (1)* $C \subseteq Clo([P]_{\mathcal{D}_{org}})$, *(2)* $C \subseteq t_+$ *and (3)* $C \supseteq P$ *(for* $P \subseteq t_+$*). According to the definition of convex space, point (1) & (3) imply that* $C \in [P]_{\mathcal{D}_{org}}$. *Combining the facts that* $C \in [P]_{\mathcal{D}_{org}}$ *and* $C \subseteq t_+$, *we have* $C \in [P]_{\mathcal{D}_{upd+}}$. *We then assume that there exists* C' *such that* $C' \supset C$ *and* $C' \in [P]_{\mathcal{D}_{upd+}}$. $C' \in [P]_{\mathcal{D}_{upd+}}$ *implies that* $C' \in [P]_{\mathcal{D}_{org}}$ *and* $C' \subseteq t_+$. $C' \in [P]_{\mathcal{D}_{org}}$ *further implies that* $C' \subseteq Clo([P]_{\mathcal{D}_{org}})$. *Then we have* $C' \subseteq Clo([P]_{\mathcal{D}_{org}})$ *and* $C' \subseteq t_+$, *and thus* $C' \subseteq C$ *(for* $C = Clo([P]_{\mathcal{D}_{org}}) \cap t_+$*). This contradicts with the initial assumption. Therefore,* $C \in [P]_{\mathcal{D}_{upd+}}$ *and there does not exist* C' *such that* $C' \supset C$ *and* $C' \in [P]_{\mathcal{D}_{upd+}}$. *According to Definition 3.4,* C *is the closed pattern of* $[P]_{\mathcal{D}_{upd+}}$.

Then we prove $Keys([P]_{\mathcal{D}_{upd+}}) = \{K | K \in Keys([P]_{\mathcal{D}_{org}}) \wedge K \subseteq t_+\}$. *First, let* $\mathcal{K} = \{K | K \in Keys([P]_{\mathcal{D}_{org}}) \wedge K \subseteq t_+\}$ *and let pattern* X *be any pattern*

that $X \in \mathcal{K}$. $X \in \mathcal{K}$ *implies that* $X \in [P]_{\mathcal{D}_{org}}$ *and* $X \subseteq t_+$. *This means* $X \in [P]_{\mathcal{D}_{upd+}}$. $X \in \mathcal{K}$ *also means* $X \in Keys([P]_{\mathcal{D}_{org}})$, *i.e.* X *is one of the most "general" patterns in* $[P]_{\mathcal{D}_{org}}$ *(Definition 3.4). Moreover,* $[P]_{\mathcal{D}_{upd+}} \subset [P]_{\mathcal{D}_{org}}$. *Therefore,* X *must also be one of the most "general" patterns in* $[P]_{\mathcal{D}_{upd+}}$. *This means that* $X \in Keys([P]_{\mathcal{D}_{upd+}})$ *for every* $X \in \mathcal{K}$. *Thus we have (A)* $\mathcal{K} \subseteq Keys([P]_{\mathcal{D}_{upd+}})$. *Second, we assume that there exists a pattern* Y *such that* $Y \in Keys([P]_{\mathcal{D}_{upd+}})$ *but* $Y \notin \mathcal{K}$. $Y \in Keys([P]_{\mathcal{D}_{upd+}})$ *means* $Y \in [P]_{\mathcal{D}_{upd+}}$. *According to the definition of* $[P]_{\mathcal{D}_{upd+}}$, *we know* $Y \in [P]_{\mathcal{D}_{org}}$ *and* $Y \subseteq t_+$. $Y \subseteq t_+$ *and* $Y \notin \mathcal{K}$ *imply that* $Y \notin Keys([P]_{\mathcal{D}_{org}})$. *This means there exists pattern* $K' \subset Y$ *that* $K' \in [P]_{\mathcal{D}_{org}}$ *(Definition 3.4). Since* $K' \subset Y$ *and* $Y \subseteq t_+$, $K' \subset t_+$, *which implies* $K' \in [P]_{\mathcal{D}_{upd+}}$. *Thus, according to Definition 3.4,* $Y \notin Keys([P]_{\mathcal{D}_{upd+}})$. *This contradicts with the initial assumption. Thus there does not exists pattern* Y *such that* $Y \in Keys([P]_{\mathcal{D}_{upd+}})$ *but* $Y \notin \mathcal{K}$. *Therefore, we have (B)* $\mathcal{K} \supseteq Keys([P]_{\mathcal{D}_{upd+}})$. *Combining results (A) and (B), we have* $Keys([P]_{\mathcal{D}_{upd+}}) = \mathcal{K} = \{K | K \in Keys([P]_{\mathcal{D}_{org}}) \wedge K \subseteq t_+\}$.

Scenario 4 is complementary to Scenario 3. The proof for the splitting of equivalence class in Scenario 4 follows exactly the same as in Scenario 3. The definitions of the closed pattern and generators for the equivalence class $[P]_{\mathcal{D}_{upd+}}$ *follows from Definition 3.4.*

Finally, we prove that Theorem 4.2 is complete. For patterns $P \in \mathcal{F}(ms_{\%}, \mathcal{D}_{org})$, *it is obvious that Scenario 1 to 4 enumerated all possible cases. For pattern* $P \notin \mathcal{F}(ms_{\%}, \mathcal{D}_{org})$, *Scenario 5 corresponds to the case where* $P \subseteq t_+$ *and* $Sup(P, \mathcal{D}_{upd+}) \geq \lceil ms_{\%} \times |\mathcal{D}_{upd+}| \rceil$. *The cases where* $P \not\subseteq t_+$ *or* $Sup(P, \mathcal{D}_{upd+}) < \lceil ms_{\%} \times |\mathcal{D}_{upd+}| \rceil$ *are not enumerated, because, in these cases, it is clear that* $P \notin \mathcal{F}(ms_{\%}, \mathcal{D}_{upd+})$. *As a result, we can conclude that Theorem 4.2 is sound and complete.*

□

Scenario 3 and 4 in Theorem 4.2 describe the cases where an existing equivalence class splits. The splitting up of an equivalence class is a bit complicated. Thus a graphical example is shown in Figure 2 (b). The original equivalence class EC_{org} splits up due to the insertion of transaction t_+. The resulting equivalence class EC'' corresponds to the equivalence class $[P]_{\mathcal{D}_{upd+}}$ described in Scenario 3, and EC' corresponds to $[P]_{\mathcal{D}_{upd+}}$ described in Scenario 4.

Theorem 4.2 summarizes how the frequent pattern space evolves when a new transaction is inserted. More importantly, the theorem describes how the updated frequent equivalence classes of $\mathcal{D}_{upd+}$ can be derived from the existing frequent equivalence classes of $\mathcal{D}_{org}$. Theorem 4.2 provides us a theoretical framework for effective incremental maintenance of the frequent pattern space. Note that: although the theorem focuses on the case where only one new transaction is inserted, it is also applicable to batch updates [3]. Suppose $\mathcal{D}_{inc} = \{t_1, \cdots, t_n\}$. To obtain the updated pattern space $\mathcal{F}(\mathcal{D}_{upd+}, ms_{\%})$, we just need to update the original space $\mathcal{F}(\mathcal{D}_{org}, ms_{\%})$ iteratively based on Theorem 4.2 for each $t_i \in \mathcal{D}_{inc}$ $(1 \leq i \leq n)$.

In addition, if the support threshold is defined in terms of percentage, $ms_{\%}$, an incremental update affects the absolute support threshold, ms_a. Recall that $ms_a = \lceil ms_{\%} \times |\mathcal{D}| \rceil$. Since $|\mathcal{D}_{upd+}| > |\mathcal{D}_{org}|$, the updated absolute support threshold $ms'_a = \lceil ms_{\%} \times |\mathcal{D}_{upd+}| \rceil \geq ms_a = \lceil ms_{\%} \times |\mathcal{D}_{org}| \rceil$. Thus, in this case, the absolute support threshold, ms_a, increases after an incremental update. Moreover, this increase in ms_a may cause some existing frequent equivalence classes to become infrequent. $EC2$ in Figure 2 (a) is an example.

Combining all the above observations, we summarize that the incremental maintenance of the frequent pattern space involves four major computational tasks: (1) update the support of existing frequent equivalence classes; (2) split

[3] A generalized version of Theorem 4.2, which describes how the frequent pattern space evolves when a batch of new transactions are added, is presented in Feng *et al.* (2009).

up equivalence classes that satisfy Scenario 3 and 4 of Theorem 4.2; (3) discover newly emerged frequent equivalence classes; and (4) remove existing frequent equivalence classes that are no longer frequent. Task (4) can be accomplished by filtering out the infrequent equivalence classes when outputting them. This filtering step is very straightforward, and thus we will not elaborate its details. We here focus on the first three tasks, and we name them respectively as the **support update** task, **class splitting** task and **new class discovery** task. To efficiently complete these three tasks, a new data structure, *Generator-Enumeration Tree* (*GE-tree*), is developed.

4.2 Maintenance Data Structure: Generator-Enumeration Tree

The *Generator-Enumeration Tree* (*GE-tree*) is a data structure inspired by the idea of the *Set-Enumeration Tree* (*SE-tree*). Thus we first recap the concept of *SE-tree*. We then introduce the characteristics of *GE-tree*, and we further demonstrate how the *GE-tree* can help to efficiently complete the computational tasks of incremental maintenance.

4.2.1 Set-Enumeration Tree

Set-Enumeration Tree (*SE-tree*), as shown in Figure 3, is a conceptual data structure that guides the systematic enumeration of patterns.

Let the set $I = \{i_1, ..., i_m\}$ of items be ordered according to an arbitrary ordering $<_0$ so that $i_1 <_0 i_2 <_0 \cdots <_0 i_m$. For itemsets X, $Y \subseteq I$, we write $X <_0 Y$ iff X is lexicographically "before" Y according to the order $<_0$. E.g. $\{i_1, i_2\} <_0 \{i_1, i_3\} <_0 \{i_1, i_2, i_3\}$. We say an itemset X is a "prefix" of an itemset Y iff $X \subseteq Y$ and $X <_0 Y$. We write $last(X)$ for the item $\alpha \in X$, if the items in X are $\alpha_1 <_0 \alpha_2 <_0 \cdots <_0 \alpha$. We say an itemset X is the "precedent"

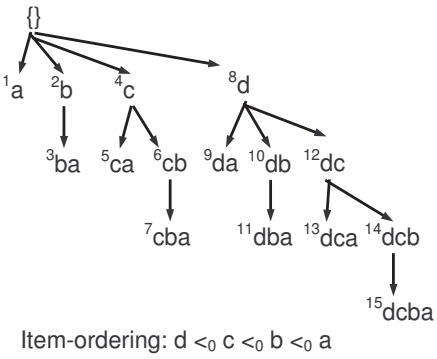


Figure 3: The Set-Enumeration Tree with item order: $d <_0 c <_0 b <_0 a$. The number on the left top corner of each node indicates the order at which the node is visited.

of an itemset Y iff $X = Y - last(Y)$. E.g. pattern $\{d, c\}$ in Figure 3 (a) is the precedent of pattern $\{d, c, b\}$.

A *SE-tree* is a conceptual organization on the subsets of I so that $\{\}$ is its root node; for each node X such that Y_1, ..., Y_k are all its children from left to right, then $Y_k <_0 \cdots <_0 Y_1$; for each node X in the set-enumeration tree such that X_1, ..., X_k are siblings to its left, we make $X \cup X_1$, ..., $X \cup X_k$ the children of X; $|X \cup X_i| = |X| + 1 = |X_i| + 1$; and $|X| = |X_i| = |X \cap X_i| + 1$. We also induce an enumeration ordering on the nodes of the *SE-tree* so that given two nodes X and Y, we say $X <_1 Y$ iff X would be visited before Y when we visit the set-enumeration tree in a left-to-right top-down manner. Since this visit order is a bit unusual, we illustrate it in Figure 3. Here, the number besides the node indicates the order at which the node is visited.

The *SE-tree* is an effective structure for pattern enumeration. Its left-to-right top-down enumeration order effectively ensures complete pattern enumeration without redundancy.

4.2.2 Generator-Enumeration Tree

The *Generator-Enumeration Tree* (*GE-tree*) is developed from the *SE-tree*. As shown in Figure 4 (a), *GE-tree* is constructed in a similar way as *SE-tree*, and *GE-tree* also follows the left-to-right top-down enumeration order to ensure complete and efficient pattern enumeration.

New features have been introduced to the *GE-tree* to facilitate incremental maintenance of frequent patterns. In the literature, *SE-tree* has been used to enumerate frequent patterns (Wang *et al.*, 2000), closed patterns (Wang *et al.*, 2003) and maximal patterns (Bayardo, 1998). However, *GE-tree*, as the name suggested, is employed here to enumerate frequent generators. Moreover, unlike *SE-tree*, in which the items are arranged according to some arbitrary order, items in *GE-tree* is arranged based on the support of the items. This means items $i_1 <_0 i_2$ if $sup(\{i_1\}, \mathcal{D}) < sup(\{i_2\}, \mathcal{D})$. This item ordering effectively minimizes the size of the *GE-tree*. Also, different from *SE-tree*, which only acts as a conceptual data structure, *GE-tree* acts as a compact storage structure for frequent generators. As shown in Figure 4, each node in *GE-tree* represents a generator, and each frequent generator is linked to its corresponding equivalence class. This feature allows frequent generators and their corresponding equivalence classes to be easily updated in the response of updates. The most important feature of *GE-tree* is that: it stores the "negative generator border" in addition to frequent generators. For the *GE-tree* in Figure 4, the "negative generator border" refers to the collection of generators under the solid line. The "negative generator border" is a newly defined concept for effective enumeration of new frequent generator and equivalence classes.

More details of these new features will be discussed as we demonstrate how *GE-tree* can help to effectively complete the computational tasks of incremental maintenance. Recall that the major computational tasks in the incremental

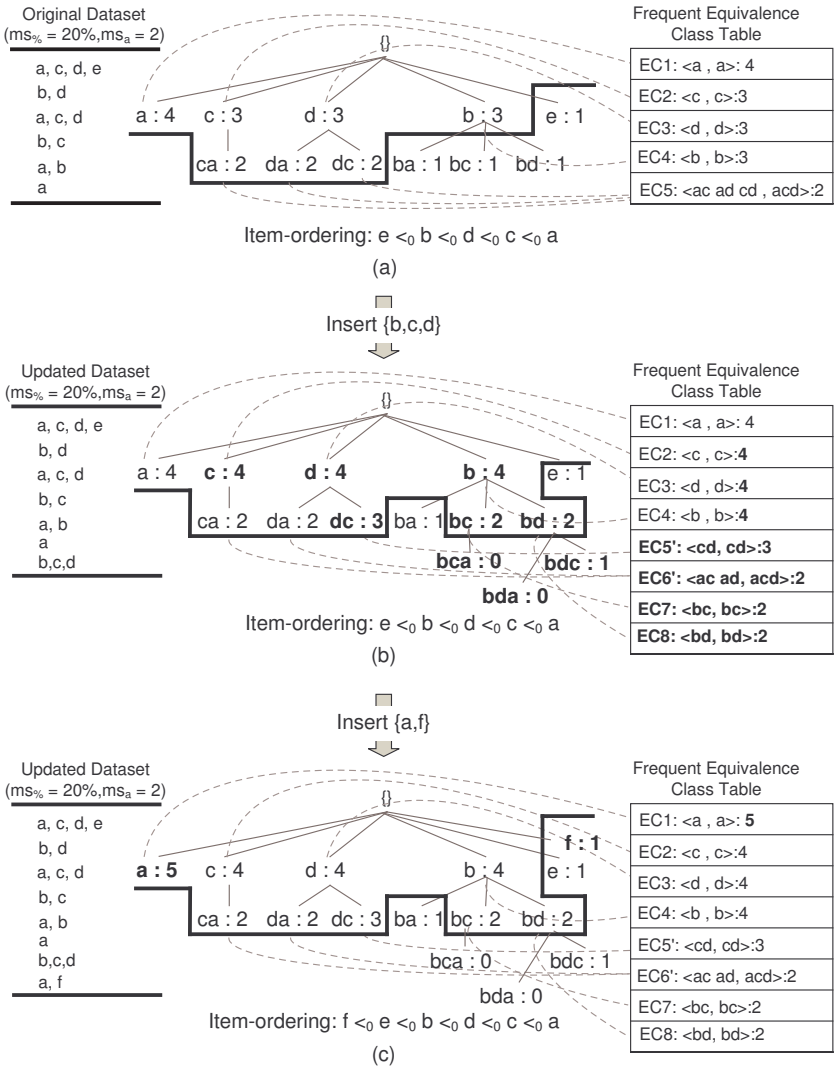


Figure 4: (a) The *GE-tree* for the original dataset. (b) The updated *GE-tree* when new transaction $\{b, c, d\}$ is inserted. (c) The updated *GE-tree* when new transaction $\{a, f\}$ is inserted.

maintenance of the frequent pattern space include the support update task, class splitting task and new class discovery task.

Support update of existing frequent equivalence classes can be efficiently accomplished with *GE-tree*. The main idea is to update only the frequent equivalence classes that need to be updated. We call these equivalence classes the "affected classes", and we need a fast way to locate these affected classes.

Since generators are the right bound of equivalence classes, finding frequent generators that need to be updated is equivalent to finding the equivalence classes. *GE-tree* can help us to locate these generators effectively. Suppose a new transaction t_+ is inserted. We will traverse the *GE-tree* in the left-to-right top-down manner. However, we usually do not need to traverse the whole tree. For any generator X in the *GE-tree*, X needs to be updated iff $X \subseteq t_+$. If $X \not\subseteq t_+$, according to Scenario 1 in Theorem 4.2, no update action is needed for X and its corresponding equivalence classes. Furthermore, according to the "a prior" property of generators (Fact 3.6), all the children of X can be skipped for the traverse. For example, in Figure 4 (c), when transaction $\{a, f\}$ is inserted, only node $\{a\}$ needs to be updated and all the other nodes are skipped.

class splitting task can also be completed efficiently with the help of *GE-tree*. The key here is to effectively locate existing frequent equivalence classes that need to be split. Extended from Scenario 3 and 4 in Theorem 4.2, we have the following corollary.

Corollary 4.3 *Suppose a new transaction t_+ is inserted into the original dataset $\mathcal{D}_{org}$. An existing frequent equivalence class $[P]_{\mathcal{D}_{org}}$ splits into two iff $P \subseteq t+$ but $Clo([P]_{\mathcal{D}_{org}}) \not\subseteq t_+$, where $Clo([P]_{\mathcal{D}_{org}})$ is the closed pattern of $[P]_{\mathcal{D}_{org}}$.*

Therefore, for an affected class X that has been identified in the support update step, X splits into two iff $Clo(X) \not\subseteq t_+$. In Figure 4, equivalence class

$EC5$ splits into two, $EC5'$ & $EC6'$, after the insertion of $\{b, c, d\}$. This is because pattern $\{c, d\}(\in EC5) \subset \{b, c, d\}$ but $Clo(EC5) = \{a, c, d\} \not\subseteq \{b, c, d\}$.

New class discovery task is the most challenging computational task involved in the incremental maintenance of the frequent pattern space. This is because, unlike the existing frequent equivalence classes, we have little information about the newly emerged frequent equivalence classes. To address this challenge, a new concept — the "negative generator border" is introduced.

4.2.3 Negative Generator Border

The "negative generator border" is defined based on the the idea of "negative border". The notion of negative border is first introduced in Mannila and Toivonen (1997). The negative border of frequent patterns refers to the set of minimal infrequent patterns. On the other hand, the negative generator border, as formally defined in Definition 4.4, refers to the set of infrequent generators that have frequent precedents in the *GE-tree*. In Figure 4, the generators under the solid line are "negative border generators", and the collection of all these generators forms the "negative generator border".

Definition 4.4 (Negative Generator Border) *Given a dataset $\mathcal{D}$, support threshold $ms_{\%}$ and the GE-tree, a pattern P is a "negative border generator" iff (1) P is a generator, (2) P is infrequent, (3) the precedent of P in the GE-tree is frequent. The set of all negative border generators is called the "negative generator border".*

As can be seen in Figure 4, the negative generator border records the nodes, where the previous enumeration stops. It thus serves as a convenient starting point for further enumeration of newly emerged frequent generators. This allows us to utilize previously obtained information to avoid redundant generation of existing generator and enumeration of unnecessary candidates.

When new transactions are inserted, the negative generator border is updated along with the frequent generators. Take Figure 4 (b) as an example. After the insertion of $\{b, c, d\}$, two negative border generators $\{b, c\}$ and $\{b, d\}$ become frequent. As a result, these two generators will be promoted as frequent generators, and their corresponding equivalence classes $EC7$ and $EC8$ will also be included into the frequent pattern space. Moreover, these two newly emerged frequent generators now act as starting pointing for further enumeration of generators. Following the *SE-tree* enumeration manner, the children of $\{b, c\}$ and $\{b, d\}$ are enumerated by combining $\{b, c\}$ and $\{b, d\}$ with their left hand side siblings, as demonstrated in Figure 4 (b). We discover that, after new transactions are added, the negative generator border expands and moves away from the root of *GE-tree*.

The detailed enumeration process is presented in Procedure 1. In Procedure 1 the following notations are used: *NG.support* denotes the support of generator *NG*; *EC.close* refers to the closed pattern of the equivalence class *EC*; *EC.keys* refers to the generators of *EC* and *GE-tree.ngb* refers to the negative generator border of the *GE-tree*.

In summary, *GE-tree* is an effective data structure that not only compactly stores the frequent generators but also guides efficient enumeration of generators. We have demonstrated with examples that the *GE-tree* greatly facilitate the incremental maintenance of the frequent pattern space.

4.3 Proposed Algorithm: PSM+

A novel incremental maintenance algorithm, *Pattern Space Maintenance+* (PSM+), is proposed based on the *GE-tree*. The pseudo-code of PSM+ is presented in Algorithm 2 and Procedure 1. In Algorithm 2 and Procedure 1, we use *X.support* to denote the support of pattern X or equivalence class X; we use

Procedure 1 enumNewEC

Input: NG, a starting point for enumeration; $\mathcal{F}$ the set of frequent equivalence classes; ms_a the absolute support threshold and *GE-tree*.
Output: $\mathcal{F}$ and the updated *GE-tree*.
Method:

```
1: if NG.support ≥ ms_a then
2:    //Newly emerged frequent generator and equivalence class.
3:    Let C be the corresponding closed pattern of NG;
4:    if ∃EC ∈ F such that EC.close = C then
5:       NG → EC.keys;
         {The corresponding equivalence class already exists.}
6:    else
7:       Create new equivalence class EC';
8:       EC'.close = C;
9:       NG → EC'.keys;
10:      EC' → F;
11:   end if
      {Enumerate new generators from NG}
12:   for all X, where X is the left hand side sibling of NG in GE-tree do
13:      NG' := NG ∪ X;
14:      if NG' is a generator then
15:         enumNewEC(NG', F, ms_a, GE-tree);
16:      end if
17:   end for
18: else
19:   NG → GE-tree.ngb; {New negative generator border.}
20: end if
21: return F and GE-tree;
```

$X.close$ to denote the closed pattern of equivalence class X and we use $X.keys$ to denote the set of generators of equivalence class X. We have also proven the correctness of PSM+.

Theorem 4.5 *PSM+ presented in Algorithm 2 correctly maintains the frequent pattern space, which is represented using equivalence classes, for incremental updates.*

Proof: *According to Theorem 4.2, after the insertion of each new transaction t_+, there are only 5 scenarios for any frequent equivalence class $[P]_{\mathcal{D}_{upd+}}$. We prove the correctness of our algorithm according to these 5 scenarios.*

For Scenario 1, suppose (i)$P \in \mathcal{F}(ms_{\%}, \mathcal{D}_{org})$, (ii) $P \not\subseteq t_+$ and (iii) $Q \not\subseteq t_+$ for all $Q \in [P]_{\mathcal{D}_{org}}$. Point (i) implies that $[P]_{\mathcal{D}_{org}}$ is an existing frequent equivalence class. Then Point (iii) implies that none of the generators of $[P]_{\mathcal{D}_{org}}$ will satisfy the condition in Line 4. As a result, $[P]_{\mathcal{D}_{org}}$ will skip all the maintenance actions and remain unchanged as desired.

For Scenario 2, suppose (i)$P \in \mathcal{F}(ms_{\%}, \mathcal{D}_{org})$, (ii) $P \subseteq t_+$ and (iii) $Q \subseteq t_+$

Algorithm 2 PSM+

Input: $\mathcal{D}_{inc}$ the incremental dataset; $|\mathcal{D}_{upd+}|$ the size of the updated dataset; $\mathcal{F}_{org}$ the original frequent pattern space represented using equivalence classes; *GE-tree* and $ms_{\%}$ the support threshold.

Output: $\mathcal{F}_{upd+}$ the update frequent pattern space represented using equivalence classes and the updated *GE-tree*.

Method:

1: $\mathcal{F} := \mathcal{F}_{org}$; {Initialization.}
2: $ms_a = \lceil ms_{\%} \times |\mathcal{D}_{upd+}| \rceil$;
3: **for all** transaction t in $\mathcal{D}_{inc}$ **do**
4: **for all** generator G in *GE-tree* that $G \subseteq t$ **do**
5: $G.support := G.support + 1$;
6: **if** G is an existing frequent generator **then**
7: Let EC be the equivalence class of G in $\mathcal{F}$;
8: **if** $EC.close \subseteq t$ **then**
9: $EC.support = G.support$;{Corresponds to Scenario 2 of Theorem 4.2.}
10: **else**
11: //split up EC {Corresponds to Scenario 3 & 4 of Theorem 4.2.}
12: $EC.keys = min\{\{K | K \in EC.keys \wedge K \not\subseteq t\} \cup \{K' \cup \{x_i\} | K' \in EC.keys \wedge K' \subseteq t, x_i \in EC.close \wedge x_i \notin t\}\}$;
13: $C = EC.close \cap t$;
14: **if** $\exists EC \in \mathcal{F}$ such that $EC.close = C$ **then**
15: $EC.support = G.support$; {EC already exists.}
16: $G \rightarrow EC.keys$;
17: **else**
18: Create new equivalence class EC';
19: $EC'.support = G.support$
20: $G \rightarrow EC'.keys$;
21: $EC' \rightarrow \mathcal{F}$;
22: **end if**
23: **end if**
24: **else if** $G.support \geq ms_a$ **then**
25: enumNewEC(NG, $\mathcal{F}$, ms_a, *GE-tree*); {Corresponds to Scenario 5 of Theorem 4.2.}
26: **end if**
27: **end for**
28: **end for**
29: Include the frequent equivalence classes in $\mathcal{F}$ into $\mathcal{F}_{upd+}$;
30: **return** $\mathcal{F}_{upd+}$ and the updated *GE-tree*;

for all $Q \in [P]_{\mathcal{D}_{org}}$. *Point (iii) implies that the generators of* $[P]_{\mathcal{D}_{org}}$ *satisfy the condition in Line 4, and the support of the generators will be updated by Line 5. Point (i) implies that* $[P]_{\mathcal{D}_{org}}$ *is an existing frequent equivalence class. Thus the generators of* $[P]_{\mathcal{D}_{org}}$ *are existing frequent generators, which satisfy the condition in Line 6. Then Point (iii) also implies that the closed pattern of* $[P]_{\mathcal{D}_{org}}$ *will satisfy the condition in Line 8. Therefore, the support of* $[P]_{\mathcal{D}_{org}}$ *will be updated in Line 9, but* $[P]_{\mathcal{D}_{org}}$ *remains unchanged as desired.*

For Scenario 3, suppose (i) $P \in \mathcal{F}(ms_{\%}, \mathcal{D}_{org})$, *(ii)* $P \subseteq t_+$ *and (iii)* $Q \not\subseteq t_+$ *for some* $Q \in [P]_{\mathcal{D}_{org}}$. *Point (ii) implies that some generators of* $[P]_{\mathcal{D}_{org}}$ *will satisfy the condition in Line 4, and Point (i) implies the condition in Line 6 is also satisfied. Then Point (iii) implies that the condition in Line 8 is not*

satisfied. Thus the equivalence class will be split into two as desired. $[P]_{\mathcal{D}_{upd+}}$ *described in Scenario 3 is updated in Line 13 to 22.*

For Scenario 4, suppose (i) $P \in \mathcal{F}(ms_{\%}, \mathcal{D}_{org})$*, (ii)* $P \not\subseteq t_+$ *and (iii)* $Q \subseteq t_+$ *for some* $Q \in [P]_{\mathcal{D}_{org}}$*. Point (iii) implies that some generators of* $[P]_{\mathcal{D}_{org}}$ *will satisfy the condition in Line 4, and Point (i) implies the condition in Line 6 is also satisfied. Then Point (ii) implies that the condition in Line 8 is not satisfied. Thus the equivalence class will be split into two as desired. Being complement to Scenario 3,* $[P]_{\mathcal{D}_{upd+}}$ *described in Scenario 4 is updated in Line 12.*

For Scenario 5, suppose (i) $P \notin \mathcal{F}(ms_{\%}, \mathcal{D}_{org})$*, (ii)* $P \subseteq t_+$ *and (iii)* $Sup(P, \mathcal{D}_{upd+}) \geq \lceil ms_{\%} \times |\mathcal{D}_{upd+}| \rceil$*. Point (ii) implies that some generators of* $[P]_{\mathcal{D}_{upd+}}$ *will satisfy the condition in Line 4. Point (i) implies that* $[P]_{\mathcal{D}_{upd+}}$ *is an existing frequent equivalence class, and thus Line 6 is not satisfied. Then we check Line 24. Point (iii) implies that the generators of* $[P]_{\mathcal{D}_{upd+}}$ *satisfy the condition in Line 24. Therefore, we will go to Line 25 and go into Procedure 1. In Line 3 to 11 of Procedure 1,* $[P]_{\mathcal{D}_{upd+}}$ *is then constructed and included as a newly emerged frequent equivalence class as desired.*

Finally, since an incremental update induces the data size and the absolute support threshold to increase, Line 29 is put in to remove equivalence classes that are no longer frequent. With that, the theorem is proven. □

We have proven that PSM+ is correct. Now we demonstrate that PSM+ is also computational effective. Recall that the incremental maintenance of frequent patterns involves three major computational tasks: the support update task, class splitting task and new class discovery task. We have demonstrated that, with the help of *GE-tree*, the support update task and the class splitting task can be efficiently completed with little computational overhead. Therefore, the major contribution to the time complexity of PSM+ comes from the new class discovery task. For the new class discovery task, the computational

dataset	#PSM+	#FPgrwoth*	#GC-growth
bms-pos ($ms_{\%} = 0.1\%$)	80	110K	110K
bms-webview1 ($ms_{\%} = 0.1\%$)	250	3K	3K
chess ($ms_{\%} = 40\%$)	350K	6M	1M
connect-4 ($ms_{\%} = 20\%$)	80K	1800M	1M
mushroom ($ms_{\%} = 0.5\%$)	10K	300M	165K
pumsb* ($ms_{\%} = 30\%$)	2K	400K	27K
retail ($ms_{\%} = 0.1\%$)	270	8K	8K
T10I4D100K ($ms_{\%} = 0.5\%$)	11	1K	1K
T40I10D100K ($ms_{\%} = 10\%$)	7K	70K	55K

Table 1: Comparison of the number of patterns enumerated by PSM+, FP-grwoth* and GC-grwoth. Notations: #PSM+, #FPgrwoth* and #GC-growth denote the approximated number of patterns enumerated by the respectively algorithms.

complexity is proportional to the number of patterns enumerated. As a result, the time complexity of PSM+ can be approximated as $O(N_{enum})$, where N_{enum} is the number of patterns enumerated. We have conducted some experiments to compare the number of patterns enumerated by PSM+ with the ones of FPgrowth* and GC-growth. FPgrowth* is one of the fastest frequent pattern discovery algorithms, and GC-growth is the fastest discovery algorithm for frequent equivalence classes. In the experiment, the number of patterns enumerated is recorded for the scenario where the size of new transactions $\mathcal{D}_{inc}$ is 10% of the original data size. The comparison results are summarized in Table 1. We observe that the number of patterns enumerated by PSM+ is smaller than the other two by a few orders of magnitude. Therefore, based on computational complexity, PSM+ is much more effective than FPgrowth* and GC-growth.

5 Decremental Maintenance of Pattern Space

This section discusses the decremental maintenance of the frequent pattern space. In the decremental update, some old transactions $\mathcal{D}_{dec}$ are removed from the original dataset $\mathcal{D}_{org}$, and thus the updated dataset $\mathcal{D}_{upd-} = \mathcal{D}_{org} - \mathcal{D}_{dec}$.

Given a support threshold $ms_{\%}$, the task of decremental maintenance is to obtain the updated pattern space by maintaining the original pattern space.

To develop effective decremental maintenance algorithm, we start off with a study on the evolution of the frequent pattern space under decremental updates using the concept of equivalence class. Through the space evolution study, we summarize the major computational tasks in the decremental maintenance. We then demonstrate how these computational tasks can also be completed efficiently using *GE-tree*. Finally, a novel decremental maintenance algorithm, named *Pattern Space Maintainer-* (PSM-), is proposed.

5.1 Evolution of Pattern Space

There is an obvious duality between incremental updates and decremental updates. In particular, if we first increment a dataset with $\mathcal{D}_{inc}$ and then decrement the result with $\mathcal{D}_{dec} = \mathcal{D}_{inc}$, we get back the original dataset. Conversely, if we first decrement a dataset with $\mathcal{D}_{dec}$ and then increment the result with $\mathcal{D}_{inc} = \mathcal{D}_{dec}$, we get back the original dataset. Therefore, the decremental maintenance is actually the reverse process of incremental maintenance.

After an incremental update, new frequent equivalence classes may emerge; in contrast, existing frequent equivalence classes may become infrequent after a decremental update. Moreover, for those existing frequent equivalence classes that are still frequent after the decremental update, they may evolve in three different ways. The first way is to remain unchanged without any change in support. The second way is to remain unchanged but with an decreased support. The third way is to merge with other classes. We know from Proposition 4.1 that an equivalence class may shrink in size and increase in support after an incremental update. It follows by duality that an equivalence class may increase in size (by merging) and decrease in support after a decremental update.

Corollary 5.1 *Let P be a pattern in $\mathcal{D}_{upd-}$. Then $[P]_{\mathcal{D}_{upd-}} \supseteq [P]_{\mathcal{D}_{org}}$, and $sup(P, \mathcal{D}_{upd-}) \leq sup(P, \mathcal{D}_{org})$.*

To have a deeper understanding on how the frequent pattern space evolves under the decremental update, we investigate the exact conditions for each evolution scenario to occur. We denote the closed pattern of an equivalence class $[p]_{\mathcal{D}}$ as $Clo([p]_{\mathcal{D}})$ and the generators or key patterns of $[p]_{\mathcal{D}}$ as $Keys([p]_{\mathcal{D}})$.

Theorem 5.2 *Let $\mathcal{D}_{org}$ be the original dataset, $\mathcal{D}_{dec}$ be the decremental dataset, $\mathcal{D}_{upd-} = \mathcal{D}_{org} - \mathcal{D}_{dec}$ and $ms_{\%}$ be the support threshold. For simplicity, we assume $\mathcal{D}_{dec}$ consists only one transaction t_-. For every frequent equivalence class $[P]_{\mathcal{D}_{org}}$ in $\mathcal{F}(ms_{\%}, \mathcal{D}_{org})$, exactly one of the 5 scenarios below holds:*

1. *$P \notin \mathcal{D}_{dec}$ and there does not exists Q such that $Q \notin [P]_{\mathcal{D}_{org}}$ but $f(Q, \mathcal{D}_{upd-}) = f(P, \mathcal{D}_{upd-})$, corresponding to the scenario where the equivalence class remains totally unchanged. In this case, $[P]_{\mathcal{D}_{upd-}} = [P]_{\mathcal{D}_{org}}$, $sup(P, \mathcal{D}_{upd-}) = sup(P, \mathcal{D}_{org})$ and $[P]_{\mathcal{D}_{upd-}} \in \mathcal{F}(\mathcal{D}_{upd-}, ms_{\%})$.*

2. *$P \notin \mathcal{D}_{dec}$ and $f(Q, \mathcal{D}_{upd-}) = f(P, \mathcal{D}_{upd-})$ for some $Q \notin [P]_{\mathcal{D}_{org}}$, corresponding to the scenario where the equivalence class of Q has to merge into the equivalence class of P. Let all such Q's be grouped into n distinct equivalence classes $[Q_1]_{\mathcal{D}_{org}}$, ..., $[Q_n]_{\mathcal{D}_{org}}$, having representatives Q_1, ..., Q_n satisfying the condition on Q. Then $[P]_{\mathcal{D}_{upd-}} = [P]_{\mathcal{D}_{org}} \cup \bigcup_i [Q_i]_{\mathcal{D}_{org}}$, $sup(P, \mathcal{D}_{upd-}) = sup(P, \mathcal{D}_{org})$, $Clo([P]_{\mathcal{D}_{upd-}}) = Clo([P]_{\mathcal{D}_{org}})$ and $Keys([P]_{\mathcal{D}_{upd-}}) = min\{K | K \in Keys([P]_{\mathcal{D}_{org}}) \vee K \in Keys([Q_i]_{\mathcal{D}_{org}}), 1 \leq i \leq n\}$. Furthermore, $[P]_{\mathcal{D}_{upd-}} \in \mathcal{F}(\mathcal{D}_{upd-}, ms_{\%})$, and $[Q_i]_{\mathcal{D}_{upd-}} = [P]_{\mathcal{D}_{upd-}}$ for $1 \leq i \leq n$.*

3. *$P \in \mathcal{D}_{dec}$ and $sup(P, \mathcal{D}_{upd-}) < \lceil ms_{\%} \times |\mathcal{D}_{upd-}| \rceil$, corresponding to the scenario where an existing frequent equivalence class becomes infrequent. In this case, $[P]_{\mathcal{D}_{org}} \notin \mathcal{F}(\mathcal{D}_{upd-}, ms_{\%})$.*

4. $P \in \mathcal{D}_{dec}$, $sup(P, \mathcal{D}_{upd-}) \geq \lceil ms_{\%} \times |\mathcal{D}_{upd-}| \rceil$ *and there does not exists* Q *such that* $Q \notin [P]_{\mathcal{D}_{org}}$ *but* $f(Q, \mathcal{D}_{upd-}) = f(P, \mathcal{D}_{upd-})$, *corresponding to the scenario where the equivalence class remains the same but with decreased support. In this case,* $[P]_{\mathcal{D}_{upd-}} = [P]_{\mathcal{D}_{org}}$, $sup(P, \mathcal{D}_{upd-}) = sup(P, \mathcal{D}_{org}) - sup(P, \mathcal{D}_{dec})$ *and* $[P]_{\mathcal{D}_{upd-}} \in \mathcal{F}(\mathcal{D}_{upd-}, ms_{\%})$.

5. $P \in \mathcal{D}_{dec}$, $sup(P, \mathcal{D}_{upd-}) \geq \lceil ms_{\%} \times |\mathcal{D}_{upd-}| \rceil$ *and* $f(Q, \mathcal{D}_{upd-}) = f(P, \mathcal{D}_{upd-})$ *for some* $Q \notin [P]_{\mathcal{D}_{org}}$, *corresponding to the scenario where the equivalence class of* P *has to merge into the equivalence class of* Q. *This scenario is complement to Scenario 2. In this case, the equivalence class, support, generators, and closed pattern of* $[P]_{\mathcal{D}_{upd-}}$ *is same as that of* $[Q]_{\mathcal{D}_{upd-}}$, *as computed in Scenario 2.*

Proof: *Scenario 1 and 3 are obvious.*

We first prove Scenario 4. Suppose (i) $P \in \mathcal{D}_{dec}$, *(ii)* $sup(P, \mathcal{D}_{upd-}) \geq \lceil ms_{\%} \times |\mathcal{D}_{upd-}| \rceil$ *and (iii) there does not exists* Q *such that* $Q \notin [P]_{\mathcal{D}_{org}}$ *but* $f(Q, \mathcal{D}_{upd-}) = f(P, \mathcal{D}_{upd-})$. *Point (ii) implies that* $[P]_{\mathcal{D}_{upd-}} \in \mathcal{F}(\mathcal{D}_{upd-}, ms_{\%})$. *According to Corollary 5.1, every member of* $[P]_{\mathcal{D}_{org}}$ *remains to be in* $[P]_{\mathcal{D}_{upd-}}$ *after the update. Moreover, point (iii) implies that* $f(Q, \mathcal{D}_{upd-}) \neq f(P, \mathcal{D}_{upd-})$ *for every pattern* $Q \notin [P]_{\mathcal{D}_{org}}$. *This means no new members will be included into* $[P]_{\mathcal{D}_{upd-}}$. *Therefore,* $[P]_{\mathcal{D}_{upd-}} = [P]_{\mathcal{D}_{org}}$ *and* $sup(P, \mathcal{D}_{upd-}) = |f(P, \mathcal{D}_{upd-})| = |f(P, \mathcal{D}_{org}) - f(P, \mathcal{D}_{dec})| = sup(P, \mathcal{D}_{org}) - sup(P, \mathcal{D}_{dec})$.

To prove Scenario 2, suppose (i) $P \notin \mathcal{D}_{dec}$ *(ii)* $f(Q, \mathcal{D}_{upd-}) = f(P, \mathcal{D}_{upd-})$ *for some* $Q \notin [P]_{\mathcal{D}_{org}}$. *Point (ii) implies that some new patterns* $Q \notin [P]_{\mathcal{D}_{org}}$ *will be included into* $[P]_{\mathcal{D}_{upd-}}$. *Moreover, for such Qs, according to Corollary 5.1,* $Q' \in [Q]_{\mathcal{D}_{upd-}}$ *for every pattern* $Q' \in [Q]_{\mathcal{D}_{org}}$. *Thus it is also true that* $Q' \in [P]_{\mathcal{D}_{upd-}}$ *for every* $Q' \in [Q]_{\mathcal{D}_{org}}$. *Therefore, we say that* $[Q]_{\mathcal{D}_{org}}$ *merge with* $[P]_{\mathcal{D}_{org}}$ *and* $[Q]_{\mathcal{D}_{upd-}} = [P]_{\mathcal{D}_{upd-}}$. *Let all such Q's be grouped into n distinct equivalence classes* $[Q_1]_{\mathcal{D}_{org}}$, ..., $[Q_n]_{\mathcal{D}_{org}}$, *having representatives* Q_1, ..., Q_n

satisfying the condition on Q. Then we have $[P]_{\mathcal{D}_{upd-}} = [P]_{\mathcal{D}_{org}} \cup \bigcup_i [Q_i]_{\mathcal{D}_{org}}$.

Point (i) implies that $f(P, \mathcal{D}_{upd-}) = f(P, \mathcal{D}_{org})$ and thus $sup(P, \mathcal{D}_{upd-}) = sup(P, \mathcal{D}_{org})$. Also since $[P]_{\mathcal{D}_{org}} \in \mathcal{F}(\mathcal{D}_{org}, ms_{\%})$, $sup(P, \mathcal{D}_{upd-}) = sup(P, \mathcal{D}_{org}) \geq \lceil ms_{\%} \times |\mathcal{D}_{org}| \rceil \geq \lceil ms_{\%} \times |\mathcal{D}_{upd-}| \rceil$. Therefore, $[P]_{\mathcal{D}_{upd-}} \in \mathcal{F}(\mathcal{D}_{upd-}, ms_{\%})$.

Next we prove $Clo([P]_{\mathcal{D}_{upd-}}) = Clo([P]_{\mathcal{D}_{org}})$. Let $C = Clo([P]_{\mathcal{D}_{org}})$ and assume that there exists pattern $C' \supset C$ that $C' \in [P]_{\mathcal{D}_{upd-}}$. Since C is the closed pattern of $[P]_{\mathcal{D}_{org}}$ and $C' \supset C$, according to Definition 3.4, we know $C' \notin [P]_{\mathcal{D}_{org}}$ and $f(C', \mathcal{D}_{org}) \neq f(P, \mathcal{D}_{org})$. Also since $P \notin \mathcal{D}_{dec}$, $C \notin \mathcal{D}_{dec}$ ($C \in [P]_{\mathcal{D}_{org}}$) and $C' \notin \mathcal{D}_{dec}$ ($C' \supset C$). Thus $f(C', \mathcal{D}_{dec}) = \emptyset$. Therefore, $f(C', \mathcal{D}_{upd-}) = f(C', \mathcal{D}_{org}) - f(C', \mathcal{D}_{dec}) = f(C', \mathcal{D}_{org}) - \emptyset = f(C', \mathcal{D}_{org})$. Combining the facts that $f(C', \mathcal{D}_{org}) \neq f(P, \mathcal{D}_{org})$ and $f(P, \mathcal{D}_{org}) = f(P, \mathcal{D}_{upd-})$, we have $f(C', \mathcal{D}_{upd-}) \neq f(P, \mathcal{D}_{upd-})$ and $C' \notin [P]_{\mathcal{D}_{upd-}}$. This contradicts with the initial assumption. Thus we can conclude that $C' \notin [P]_{\mathcal{D}_{upd-}}$ for all $C' \supset C$. According to Fact 3.5, C is the closed pattern of $[P]_{\mathcal{D}_{upd-}}$.

Then we prove $Keys([P]_{\mathcal{D}_{upd-}}) = min\{K | K \in Keys([P]_{\mathcal{D}_{org}}) \vee K \in Keys([Q_i]_{\mathcal{D}_{org}}), 1 \leq i \leq n\}$. This formula states that the generators of the equivalence class $[P]_{\mathcal{D}_{upd-}}$ are the set of minimum (equivalent to the most general) generators in the merging equivalence classes. This basically follows from the definition of generators in Definition 3.4.

Scenario 5 is complement of Scenario 2. Therefore, it can be proven in the same way as Scenario 2.

Last we prove that the theorem is complete. For patterns $P \notin \mathcal{D}_{dec}$, it is obvious that Scenario 1 and 2 enumerated all possible cases. For patterns $P \in \mathcal{D}_{dec}$, it is also obvious that Scenario 3 to 5 enumerated all possible cases. Therefore, the theorem is complete and correct.

□

Theorem 5.2 summarizes how the frequent pattern space evolves after a decremental update. The theorem also describes how the updated frequent equivalence classes in $\mathcal{D}_{upd-}$ can be derived from the existing frequent equivalence classes of $\mathcal{D}_{org}$. Similar to Theorem 4.2, Theorem 5.2 lays a theoretical foundation for the development of effective decremental maintenance algorithms.

In addition, opposite to the incremental update, the decremental update decreases the absolute support threshold if the support threshold is initially defined in terms of percentage. Let the original absolute support $ms_a = \lceil ms_\% \times |\mathcal{D}_{org}| \rceil$. Since $|\mathcal{D}_{upd-}| = |\mathcal{D}_{org}| - |\mathcal{D}_{dec}|$, the updated absolute support threshold $ms'_a = \lceil ms_\% \times |\mathcal{D}_{upd-}| \rceil < ms_a$. This decrease in the absolute support threshold induces new frequent equivalence classes to emerge.

Combining all the above observations, we summarize that the decremental maintenance of the frequent pattern space involves four computational tasks: (1) update the support of existing frequent equivalence classes; (2) merge equivalence classes that satisfy Scenario 2 and 5 of Theorem 5.2; (3) discover newly emerged frequent equivalence classes; and (4) remove existing frequent equivalence classes that are no longer frequent. Task (4) is excluded from our discussion, for its solution is straightforward. We here focus on the first three tasks, and we name them respectively as the **support update** task, **class merging** task and **new class discovery** task.

5.2 Maintenance of Pattern Space

We investigate here how the major computational tasks in decremental maintenance of the frequent pattern space can be efficiently accomplished.

Due to the duality between the incremental and decremental maintenance, most of the computational tasks in decremental maintenance can be effectively handled with the *GE-tree*. In particular, the **support update** task in decre-

mental maintenance is actually the reverse operation of the one in incremental maintenance. Therefore, the support of existing frequent equivalence classes can be updated using *GE-tree* in the same manner described in Section 4.2.2. Except that, in decremental maintenance, the support is decremented.

For the **new class discovery** task, newly emerged frequent equivalence classes and generators can also be effectively enumerated based on the concept of negative generator border. Details of the enumeration method is presented in Procedure 1 in Section 4.2.3. Same as in incremental maintenance, the negative generator border is updated after the removal of each old transactions. However, different from incremental updates, when old transactions are removed, the negative generator border shrinks and move towards the root of *GE-tree*.

On the other hand, the **class merging** task can not be handled in the same way as the class splitting task in incremental maintenance. However, extended from the Scenario 2 in Theorem 5.2, we have the following corollary.

Corollary 5.3 *Let* $[P]_{\mathcal{D}_{org}}$ *and* $[Q]_{\mathcal{D}_{org}}$ *be two equivalence classes in* $\mathcal{D}_{org}$ *such that* $[P]_{\mathcal{D}_{org}} \cap [Q]_{\mathcal{D}_{org}} = \emptyset$, $P \notin \mathcal{D}_{dec}$ *but* $Q \in \mathcal{D}_{dec}$. *Then* $f(P, \mathcal{D}_{upd-}) = f(Q, \mathcal{D}_{upd-})$, *meaning* $[P]_{\mathcal{D}_{org}}$ *merges with* $[Q]_{\mathcal{D}_{org}}$ *in* $\mathcal{D}_{upd-}$, *iff (1)* $sup(P, \mathcal{D}_{upd-}) = sup(Q, \mathcal{D}_{upd-})$ *and (2)* $Clo([P]_{\mathcal{D}_{org}}) \supset Clo([Q]_{\mathcal{D}_{org}})$. *Here* $Clo(X)$ *denotes the closed pattern of equivalence class* X.

Proof: *We first prove the left-to-right direction. Suppose (i)* $P \notin \mathcal{D}_{dec}$, *(ii)* $Q \in \mathcal{D}_{dec}$ *and (iii)* $f(P, \mathcal{D}_{upd-}) = f(Q, \mathcal{D}_{upd-})$. *Point (ii) implies that* $sup(P, \mathcal{D}_{upd-}) = sup(Q, \mathcal{D}_{upd-})$. *Combining Point (i),(ii) and (iii), we have* $f(P, \mathcal{D}_{org}) = f(P, \mathcal{D}_{upd-}) = f(Q, \mathcal{D}_{upd-}) = f(Q, \mathcal{D}_{org}) - f(Q, \mathcal{D}_{dec})$. *This implies that* $f(P, \mathcal{D}_{org}) \subset f(Q, \mathcal{D}_{org})$. *Therefore,* $Clo([P]_{\mathcal{D}_{org}}) \supset Clo([Q]_{\mathcal{D}_{org}})$.

We then prove the right-to-left direction. Suppose (i) $sup(P, \mathcal{D}_{upd-}) = sup(Q, \mathcal{D}_{upd-})$ *and (ii)* $Clo([P]_{\mathcal{D}_{org}}) \supset Clo([Q]_{\mathcal{D}_{org}})$. *Point (ii) implies that* $f(P, \mathcal{D}_{org}) \subset f(Q, \mathcal{D}_{org})$. *Since* $P \notin \mathcal{D}_{dec}$, *we have* $f(P, \mathcal{D}_{org}) = f(P, \mathcal{D}_{upd-}) \subset$

$f(Q, \mathcal{D}_{org})$. *Combining this with Point (i), we have* $f(P, \mathcal{D}_{upd-}) = f(Q, \mathcal{D}_{upd-})$ *as desired. The corollary is proven.* □

Corollary 5.3 provides us a means to determine which two equivalence classes need to be merged after an decremental update. Based on Corollary 5.3, one way to handle the class merging task effectively is to first group the equivalence classes based on their support. This can be done efficiently using a hash table with support values as hash keys. Then, within the group of equivalence classes that shared the same support, we further compare their closed patterns. two equivalence classes are to be merged together, if their closed patterns are superset and subset to each other. Details of this merging process is presented in Algorithm 3.

5.3 Proposed Algorithm: PSM-

A novel algorithm, *Pattern Space Maintenance-* (PSM-), is proposed for the decremental maintenance of the frequent pattern space. The pseudo-code of PSM- is presented in Algorithm 3 and Procedure 1. In Algorithm 3 and Procedure 1, we use $X.support$ to denote the support of pattern X or equivalence class X; we use $X.close$ to denote the closed pattern of equivalence class X and we use $X.keys$ to denote the set of generators of equivalence class X. We have also proven the correctness of PSM-.

Theorem 5.4 *PSM- presented in Algorithm 3 correctly maintains the frequent pattern space, which is represented using equivalence classes, for decremental updates.*

Proof: *According to Theorem 5.2, after an decremental update, an existing frequent equivalence class* $[P]_{\mathcal{D}_{org}}$ *may evolve in only 5 scenarios. We prove the correctness of our algorithm according to these 5 scenarios.*

Algorithm 3 PSM-

Input: $\mathcal{D}_{dec}$ the decremental dataset; $|\mathcal{D}_{upd-}|$ the size of the updated dataset; $\mathcal{F}_{org}$ the original frequent pattern space represented using equivalence classes ; *GE-tree* and $ms_{\%}$ the support threshold.

Output: $\mathcal{F}_{upd-}$ the updated frequent pattern space represented using equivalence classes and the updated *GE-tree*.

Method:

1: $\mathcal{F} := \mathcal{F}_{org}$; {Initialization.}
2: $ms_a = \lceil ms_{\%} \times |\mathcal{D}_{upd-}| \rceil$;
3: **for all** transaction t in $\mathcal{D}_{dec}$ **do**
4: **for all** generator G in *GE-tree* that $G \subseteq t$ **do**
5: $G.support := G.support - 1$;
6: **if** G is an existing frequent generator **then**
7: Let EC be the equivalence class of G in $\mathcal{F}$;
 {Update the support of existing frequent equivalence classes.}
8: $EC.support := G.support$;
9: **end if**
10: **if** $G.support < ms_a$ **then**
11: $G \rightarrow$ *GE-tree.ngb*; {Update the negative generator border.}
12: Remove all children of G from *GE-tree.ngb*;
13: **end if**
14: **end for**
15: **end for**
16: **for all** $NG \in$ *GE-tree.ngb* that $NG.support \geq ms_a$ **do**
17: enumNewEC(NG, $\mathcal{F}$, ms_a, *GE-tree*); {Enumerate new frequent equivalence classes.}
18: **end for**
19: **for all** equivalence class $EC \in \mathcal{F}$ **do**
20: **if** $EC.support \geq ms_a$ **then**
21: **if** $\exists EC'$ that $EC'.support = EC.support$ and $EC'.close \subset EC.close$ **then**
22: **for all** EC' that $EC'.support = EC.support$ and $EC'.close \subset EC.close$ **do**
23: $EC.keys = min\{K | K \in EC.keys \wedge K \in EC'.keys\}$;
 {Merging of equivalence classes.}
24: Remove EC' from $\mathcal{F}$;
25: **end for**
26: **end if**
27: $EC \rightarrow \mathcal{F}_{upd-}$;
28: **end if**
29: **end for**
30: **return** $\mathcal{F}_{upd-}$ and the updated *GE-tree*;

For Scenario 1, suppose (i) $P \notin \mathcal{D}_{dec}$ and (ii) there does not exists Q such that $Q \notin [P]_{\mathcal{D}_{org}}$ but $f(Q, \mathcal{D}_{upd-}) = f(P, \mathcal{D}_{upd-})$. In Line 1, $[P]_{\mathcal{D}_{org}}$ is included into $\mathcal{F}$ as initialization. Then Point (i) implies that the condition in Line 4 will not be satisfied for all transactions in $\mathcal{D}_{dec}$. Thus, Line 5 to 15 will be skipped, and the support of $[P]_{\mathcal{D}_{org}}$ remains unchanged as desired. Also since $[P]_{\mathcal{D}_{org}} \in \mathcal{F}(\mathcal{D}_{org}, ms_{\%})$, $sup(P, \mathcal{D}_{upd-}) = sup(P, \mathcal{D}_{org}) \geq \lceil ms_{\%} \times |\mathcal{D}_{org}| \rceil \geq \lceil ms_{\%} \times |\mathcal{D}_{upd-}| \rceil$. Therefore, the condition in Line 20 is satisfied. Point (ii) implies that Line 21 can not be true (Corollary 5.3). As a result, $[P]_{\mathcal{D}_{org}}$ is included in $\mathcal{F}_{upd-}$ unchanged in Line 27 as desired.

For Scenario 2, suppose (i) $P \notin \mathcal{D}_{dec}$ and (ii) $f(Q, \mathcal{D}_{upd-}) = f(P, \mathcal{D}_{upd-})$

for some $Q \notin [P]_{\mathcal{D}_{org}}$. In Line 1, $[P]_{\mathcal{D}_{org}}$ is included into $\mathcal{F}$ as initialization. Same as in Scenario 1, because of Point (i), the condition in Line 4 is not satisfied, and thus Line 5 to 15 are skipped. The support of $[P]_{\mathcal{D}_{org}}$ remains unchanged as desired. With the same reasoning in Scenario 1, Line 20 will be true. Now Point (ii) implies that Line 21 is also true (Corollary 5.3). As a result, $[P]_{\mathcal{D}_{org}}$ will be merged with other equivalence classes to form $[P]_{\mathcal{D}_{upd-}}$ as desired. Finally, $[P]_{\mathcal{D}_{upd-}}$ is included in $\mathcal{F}_{upd-}$ in Line 27 as desired.

For Scenario 3, suppose (i) $P \in \mathcal{D}_{dec}$ and (ii) $sup(P, \mathcal{D}_{upd-}) < \lceil ms_{\%} \times |\mathcal{D}_{upd-}| \rceil$. As usual, $[P]_{\mathcal{D}_{org}}$ is included into $\mathcal{F}$ as initialization. Point (ii) implies that Line 20 will not be true. Therefore, $[P]_{\mathcal{D}_{org}}$ will not be included in $\mathcal{F}_{upd-}$ as desired.

For Scenario 4, suppose (i) $P \in \mathcal{D}_{dec}$, (ii) $sup(P, \mathcal{D}_{upd-}) \geq \lceil ms_{\%} \times |\mathcal{D}_{upd-}| \rceil$ and (iii) there does not exists Q such that $Q \notin [P]_{\mathcal{D}_{org}}$ but $f(Q, \mathcal{D}_{upd-}) = f(P, \mathcal{D}_{upd-})$. As usual, $[P]_{\mathcal{D}_{org}}$ is included into $\mathcal{F}$ as initialization. Point (i) implies that the condition in Line 4 will be satisfied for some transactions in $\mathcal{D}_{dec}$. Thus the support of $[P]_{\mathcal{D}_{org}}$ will be updated as desired by Line 8. Point (ii) then implies that Line 10 is not true, and thus Line 11 to 12 are skipped. Point (ii) and (iii) also implies that Line 20 will be true but Line 21 will not be true (Corollary 5.3). As a result, $[P]_{\mathcal{D}_{org}}$ will be include in $\mathcal{F}_{upd-}$ with a updated support as desired.

For Scenario 5, since it is complement to Scenario 2, patterns of Scenario 5 will also be correctly updated as explained for Scenario 2.

Finally, since an decremental update causes the data size and the absolute support threshold to drop, new frequent equivalence classes may emerge. In PSM-, all the newly emerged frequent equivalence classes will be enumerated from the negative generator border by Line 17. With that, the theorem is proven. □

Similar to PSM+, the major contribution to the time complexity of PSM-

comes from the new class discovery task. For the new class discovery task, the computational complexity is proportional to the number of patterns enumerated. As a result, the time complexity of PSM- can also be approximated as $O(N_{enum})$, where N_{enum} is the number of patterns enumerated. Moreover, the number of patterns need to be enumerated is proportional to the number of newly emerged frequent equivalence classes. In general, under decremental updates, the number of newly emerged frequent equivalence classes is much smaller than the total number of frequent equivalence classes. This theoretically demonstrates that maintaining the frequent pattern space with PSM- is definitely much more effective than re-discovering the pattern space.

6 Pattern Space Maintainer (PSM)

We have proposed a novel algorithm, PSM+, to address the incremental maintenance of the frequent pattern space, and we have also proposed a novel algorithm, PSM-, for the decremental maintenance. Although these two maintenance algorithms are discussed separately, PSM+ and PSM- share many similarities and are both developed based on the same data structure — the *GE-tree*. Thus the integration of PSM+ and PSM- involves negligible overheads. We name the integrated version of PSM+ and PSM- the *Pattern Space Maintainer*, in short PSM.

PSM is not only a useful tool for incremental and decremental maintenance, it can also be employed to maintain the space of frequent patterns for support threshold adjustment. Support threshold adjustment is a common interactive mining operation, which is used to obtain the appropriate set of frequent patterns. When the support threshold is adjusted up, existing frequent patterns and equivalence classes may become infrequent. The maintenance for this scenario is very straightforward, and thus we will not discuss it here. On the other

hand, when the support threshold is adjusted down, new (unknown) frequent patterns and equivalence classes may emerge. The maintenance for this scenario is much more challenging, for we have little information on the newly emerged patterns. In this case, PSM can be used to effectively enumerate the newly emerged equivalence classes based on the concepts of *GE-tree* and negative generator border. The detailed enumeration method is described in Procedure 1 in Section 4.2.3.

7 Experimental Studies

The computational effectiveness of the proposed algorithms is tested on the benchmark datasets from the *FIMI* Repository (`http://fimi.cs.helsinki.fi`). The performance of the proposed algorithms are compared with the state-of-the-art approaches, which includes: FPgrowth* (Grahne and Zhu, 2005), one of the fastest frequent pattern discovery algorithms, GC-grwoth (Li *et al.*, 2005), the fastest discovery algorithm for frequent equivalence classes, CanTree (Leung *et al.*, 2007), a prefix-tree based maintenance algorithm, moment (Chi *et al.*, 2006), a currently proposed algorithm that maintains frequent closed patterns and ZIGZAG (Veloso *et al.*, 2002), a frequent maximal pattern maintenance algorithm. All the experiments are run on a PC with Duo 2.4 GHz processors and 3.2 GB RAM.

Incremental Maintenance

In real applications, the size of the incremental dataset $\mathcal{D}_{inc}$ is usually much smaller than the size of the original dataset $\mathcal{D}_{org}$, e.g. a daily sales data vs. an annual sales data, an hourly stock transaction vs. a daily transaction, etc. As a result, the performance of PSM+ is evaluated for $\Delta^+ \leq 10\%$, where $\Delta^+ = |\mathcal{D}_{inc}|/|\mathcal{D}_{org}|$.

Figure 5 compares the performance of PSM+ with the discovery algorithms,

GC-growth and FPgrowth*. It is obvious that PSM+ is much more effective. In the best case, PSM+ outperforms both discovery algorithms by three orders of magnitude; in the worse case, PSM+ is still at least twice faster; on average, PSM+ outperforms the discovery algorithms by more than an order of magnitude.

Figure 6 compares the performance of PSM+ with the maintenance algorithms, CanTree, moment and ZIGZAG. It is observed that the processing time of moment increases dramatically as the update size increases. This is because moment is proposed for the special case where each time only one transaction is added, and it works based on the hypothesis that there are only minimum changes to the frequent closed patterns given such a small amount of update. Therefore, as observed, its performance degrades significantly as the update size increases. Compared with moment, PSM+ is on average three orders of magnitude faster. For some cases, PSM+ outperforms moment by up to six orders of magnitude. PSM+ outperforms ZIGZAG on average by two orders of magnitude. Among three maintenance algorithms, CanTree is the most competitive method. Therefore, let us have a more detailed comparison between PSM+ and CanTree. Table 2 presents the speed gain achieved by PSM+ compared with CanTree. It can been seen that, although the performance of CanTree looks pretty close with the one of PSM+ in Figure 6, PSM+ is at least three times faster than CanTree. Moreover, PSM+, on average, outperforms CanTree by more than an order of magnitude.

Decremental Maintenance

With the similar reason of incremental maintenance, the performance of PSM- is evaluated for $\Delta^- \leq 10\%$, where $\Delta^- = |\mathcal{D}_{dec}|/|\mathcal{D}_{org}|$. The performance of PSM- is also compared with both pattern discovery and pattern maintenance algorithms, as shown in Figure 7.

Speed Gain	mushroom ($ms_{\%} = 0.05\%$)	bms-webview1 ($ms_{\%} = 0.1\%$)	bms-pos ($ms_{\%} = 0.1\%$)
max.	3500	200	51
ave.	1700	43	16
min.	45	3	3
Speed Gain	chess ($ms_{\%} = 40\%$)	T10I4D100K ($ms_{\%} = 0.1\%$)	T40I10D100K ($ms_{\%} = 10\%$)
max.	112	1500	2300
ave.	32	247	860
min.	3	4	11

Table 2: The speed gain achieved by PSM+ compared with CanTree.

As can be seen in Figure 7 (a), compared with the pattern discovery algorithms, PSM- is at least an order of magnitude faster. According to Figure 7 (b), PSM- also outperforms ZIGZAG by more than an order of magnitude. Moreover, similar to the results of incremental maintenance, we also observe from Figure 7 (b) that, the advantage of PSM- over moment gets larger as the update size increases.

Support Adjustment Maintenance

We have also evaluated the performance of PSM for support threshold adjustment. The effectiveness of PSM is tested with various degrees of threshold adjustment. The experimental results are presented in Figure 8. As can been seen from Figure 8, PSM outperforms both the pattern discovery and pattern maintenance algorithms considerably.

Over three different types of updates, we have one common observation. We observe that the advantage of the proposed algorithms diminishes as the size (or degree) of update increases. This is because large update size or large variation in support threshold logically leads to more dramatic changes to the frequent pattern space and makes the pattern space computational more expensive to be maintained. It is inevitable that when the amount of update increases to a certain extent, the changes induced to the pattern space become so significant

that it becomes more efficient to re-discover the pattern space than to maintain and update it.

8 Conclusion

This paper has studied the incremental and decremental maintenance of the frequent pattern space. To develop efficient maintenance algorithms, we started off by analyzing how the space of frequent patterns evolves under incremental and decremental updates. Based on this space evolution analysis, we have summarized the major computation tasks involved in frequent pattern maintenance. To effectively address the maintenance computational tasks, a new data structure, *Generator-Enumeration Tree* (*GE-tree*), is developed. Based on *GE-tree*, we proposed two novel algorithms, *Pattern Space Maintainer+* (PSM+) and *Pattern Space Maintainer-* (PSM-), for the incremental and decremental maintenance of frequent patterns. We further demonstrated that PSM+ and PSM- can be easily integrated and extended to update the frequent pattern space for support threshold adjustment. We have also evaluated the effectiveness of our proposed algorithms with extensive experimental studies. Experimental results show that the proposed algorithms outperform the state-of-the-art approaches considerably.

This paper studied the evolution of the frequent pattern space. In the future, we plan to explore the evolution and maintenance of other types of pattern spaces, e.g. the space of emerging patterns, odds ratio patterns, etc.

Acknowledgements

Thanks to Yun Chi from NEC Laboratories America, Inc. for the source code of moment. Thanks to Mohammed Javeed Zaki from Rensselaer Polytechnic

Institute, USA, for the source code of ZIGZAG. This work was supported in part by an A*STAR AGS scholarship and A*STAR SERC PSF grant 072 101 0016.

References

Agrawal,R. and Imielinski,T. (1993) Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data* pp. 207–216.

Bayardo,R.J. (1998) Efficiently mining long patterns from databases. In *Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data* pp. 85–93.

Cheung,D.W.L., Han,J., Ng,V.T.Y. and Wong,C.Y. (1996) Maintenance of discovered association rules in large databases: an incremental updating technique. In *Proceedings of the Twelfth International Conference On Data Engineering* pp. 106–114.

Cheung,D.W.L., Lee,S.D. and Kao,B. (1997) A general incremental technique for maintaining discovered association rules. In *Proceedings of the Fifth International Conference on Database Systems for Advanced Applications* pp. 185–194.

Chi,Y., Wang,H., Yu,P.S. and Muntz,R.R. (2006) Catch the moment: maintaining closed frequent itemsets over a data stream sliding window. *Knowledge and Information Systems,* **10** (3), 265–294.

Feng,M., Dong,G., Li,J., Tan,Y.P. and Wong,L. (2009) Evolution of frequent pattern space. *Information Processing Letters,* **n.a.** (n.a.), in press.

Gouda,K. and Zaki,M.J. (2001) Efficiently mining maximal frequent itemsets. In *Proceedings of the 2001 IEEE International Conference on Data Mining* pp. 163–170.

Grahne,G. and Zhu,J. (2005) Fast algorithms for frequent itemset mining using fp-trees. *IEEE Transactions on Knowledge and Data Engineering,* **17** (10), 1347–1362.

Han,J., Pei,J. and Yin,Y. (2000) Mining frequent patterns without candidate generation. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data* pp. 1–12.

Lee,C.H., Lin,C.R. and Chen,M.S. (2005) Sliding window filtering: an efficient method for incremental mining on a time-variant database. *Information Systems,* **30** (3), 227–244.

Leung,C.K.S., Khan,Q.I., Li,Z. and Hoque,T. (2007) Cantree: a canonical-order tree for incremental frequent-pattern mining. *Knowledge and Information Systems,* **11** (3), 287–311.

Li,C., Cong,G., Tung,A.K.H. and Wang,S. (2004) Incremental maintenance of quotient cube for median. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* pp. 226–235.

Li,H., Li,J., Wong,L., Feng,M. and Tan,Y.P. (2005) Relative risk and odds ratio: a data mining perspective. In *Proceedings of the Twenty-fourth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems* pp. 368–377.

Mannila,H. and Toivonen,H. (1997) Levelwise search and borders of theories in knowledge discovery. *Data Mining and Knowledge Discovery,* **1** (3), 241–258.

Pasquier,N., Bastide,Y., Taouil,R. and Lakhal,L. (1999) Discovering frequent closed itemsets for association rules. In *Proceedings of Seventh International Conference of Data Theories* pp. 398–416.

Rymon,R. (1992) Search through systematic set enumeration. In *Principles of Knowledge Representation and Reasoning* pp. 539–550.

Veloso,A., Jr.,W.M., de Carvalho,M., Pôssas,B., Parthasarathy,S. and Zaki,M.J. (2002) Mining frequent itemsets in evolving databases. In *Proceedings of the Second SIAM International Conference on Data Mining.*

Wang,J., Han,J. and Pei,J. (2003) Closet+: searching for the best strategies for mining frequent closed itemsets. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* pp. 236–245.

Wang,K., He,Y. and Han,J. (2000) Mining frequent itemsets using support constraints. In *Proceedings of 26th International Conference on Very Large Data Bases* pp. 43–52.

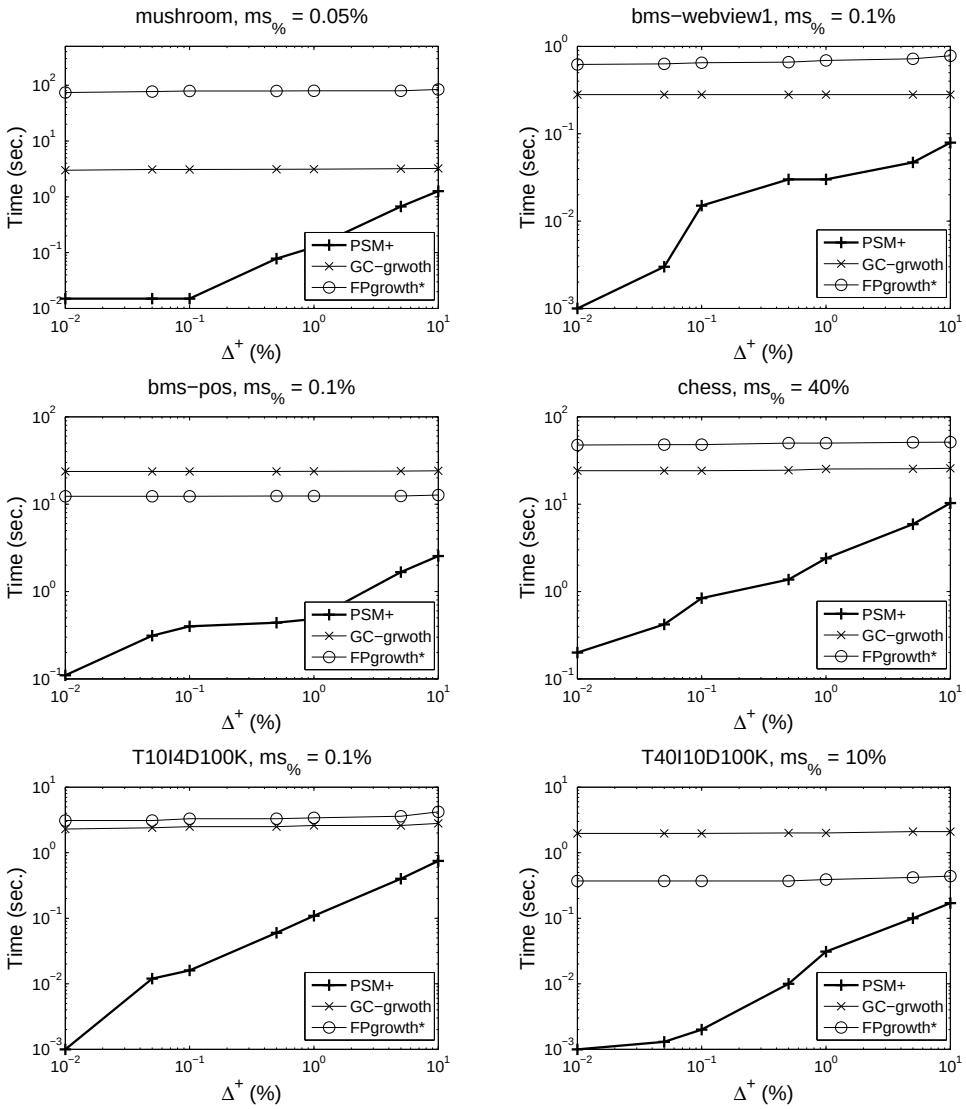


Figure 5: Performance comparison of PSM+ and the pattern discovery algorithms: GC-growth and FPgrowth* . Notations: $\Delta^+ = |\mathcal{D}_{inc}|/|\mathcal{D}_{org}|$.

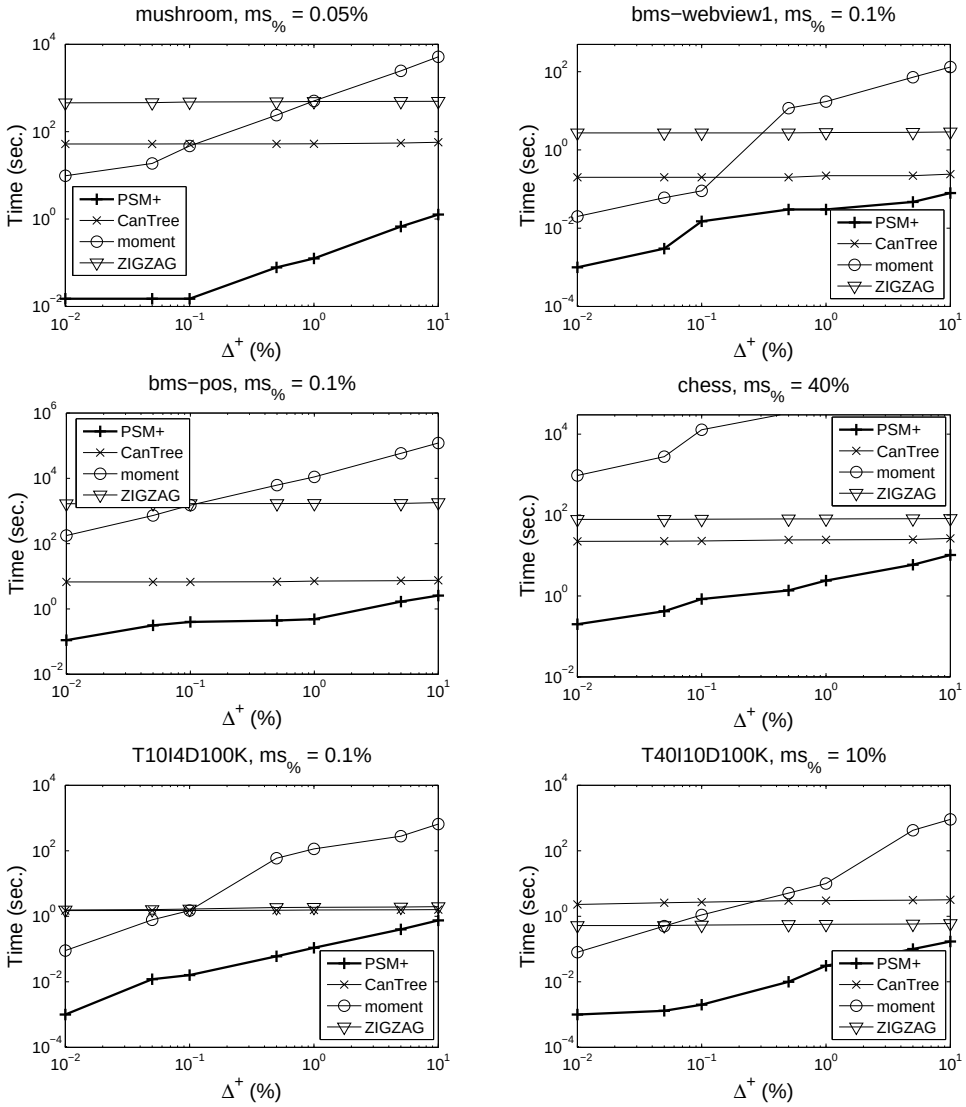
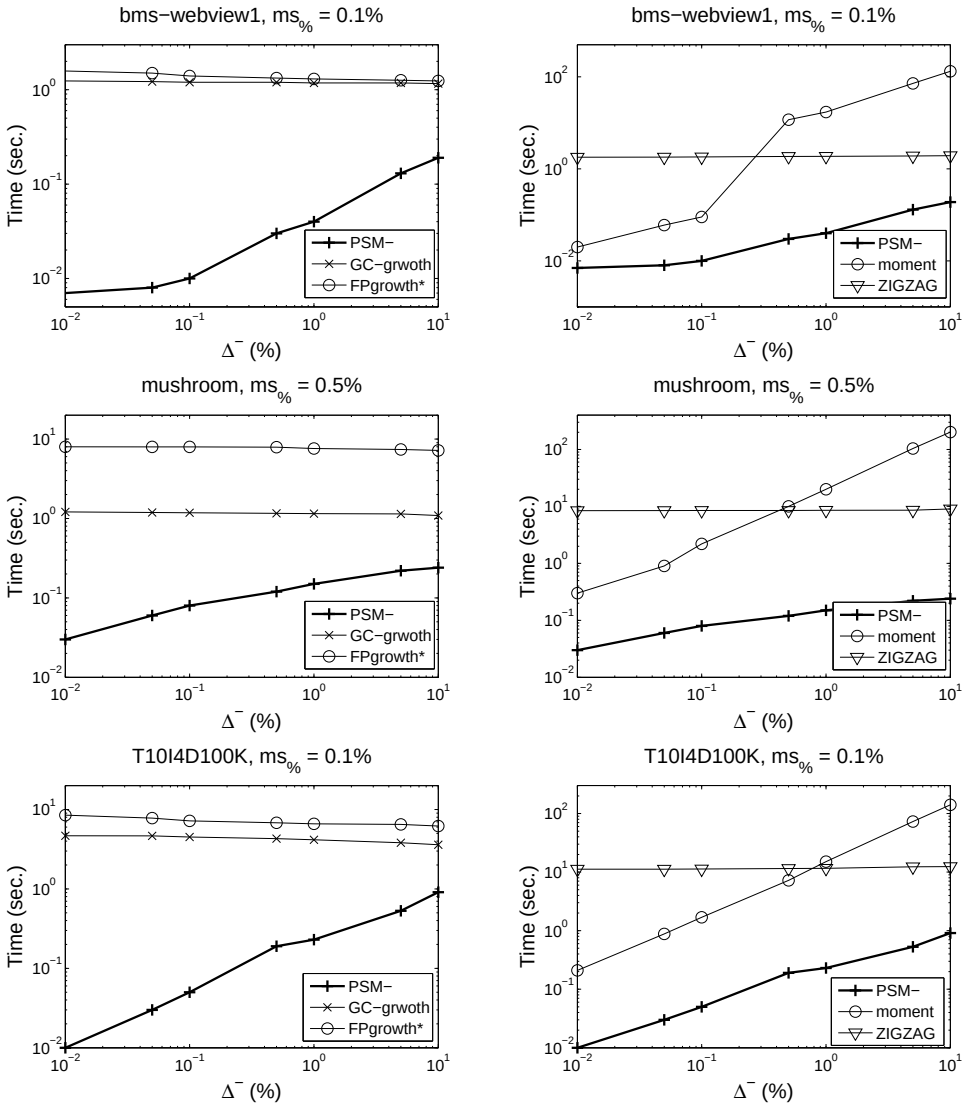


Figure 6: Performance comparison of PSM+ and the pattern maintenance algorithms, CanTree, moment and ZIGZAG. Notations: $\Delta^+ = |\mathcal{D}_{inc}|/|\mathcal{D}_{org}|$.



(a) (b)

Figure 7: (a) Performance comparison of PSM+ and the pattern discovery algorithms: GC-growth and FPgrowth*. (b)Performance comparison of PSM+ and the pattern maintenance algorithms: moment and ZIGZAG. Notations: $\Delta^- = |\mathcal{D}_{dec}|/|\mathcal{D}_{org}|$.

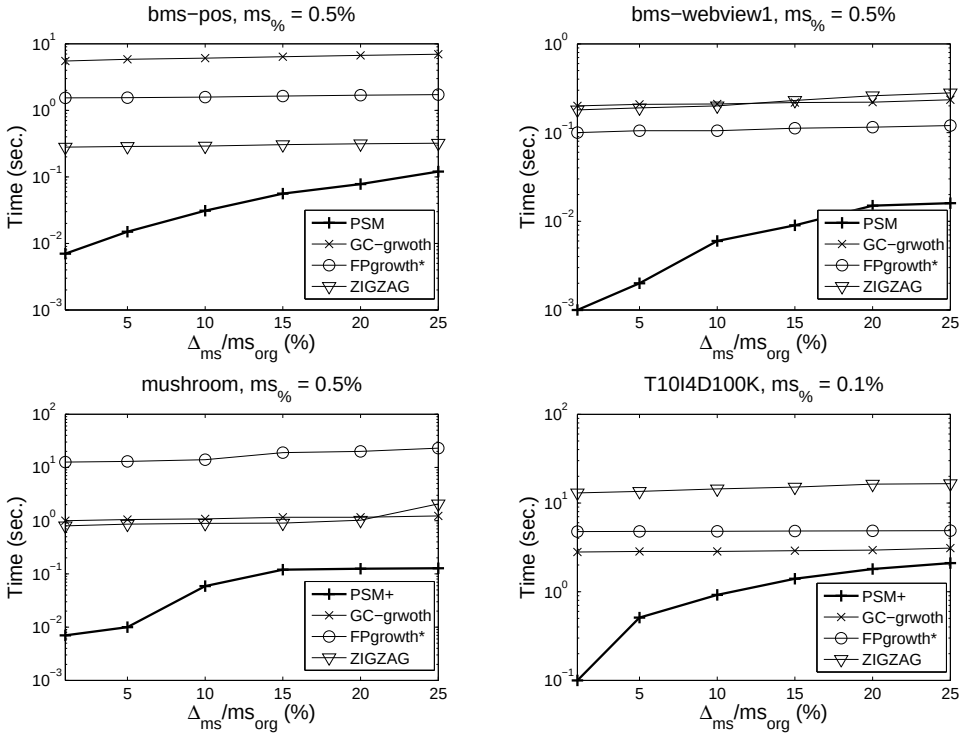


Figure 8: Performance of PSM on support threshold adjustment maintenance. Notations: Δ_{ms} denotes the difference between the original support threshold and the updated support threshold.