Research Article

Jiang Liu*, Jing Li, Feng Ni, Xiang Xia, Shunlong Li, and Wenhui Dong

An algebraic semigroup method for discovering maximal frequent itemsets

https://doi.org/10.1515/math-2022-0516
received October 10, 2021; accepted September 26, 2022

Abstract: Discovering maximal frequent itemsets is an important issue and key technique in many data mining problems such as association rule mining. In the literature, generating maximal frequent itemsets proves either to be NP-hard or to have $O(l^3 4^l(m + n))$ complexity in the worst case from the perspective of generating maximal complete bipartite graphs of a bipartite graph, where m, n are the item number and the transaction number, respectively, and l denotes the maximum of $|C||\Psi(C)|/(|C| + |\Psi(C)| - 1)$, with the maximum taken over all maximal frequent itemsets C. In this article, we put forward a method for discovering maximal frequent itemsets, whose complexity is $O(3mn2^{\beta} + 4^{\beta}n)$, lower than the known complexity both in the worst case, from the perspective of semigroup algebra, where β is the number of items whose support is more than the minimum support threshold. Experiments also show that an algorithm based on the algebraic method performs better than the other three well-known algorithms. Meanwhile, we explore some algebraic properties with respect to items and transactions, prove that the maximal frequent itemsets are exactly the simplified generators of frequent itemsets, give a necessary and sufficient condition for a maximal $i + 1$-frequent itemset being a subset of a closed i-frequent itemset, and provide a recurrence formula of maximal frequent itemsets.

Keywords: maximal frequent itemset, association rule mining, generator, semigroup

MSC 2020: 20-08, 68W99

1 Introduction

In 1993, Agrawal et al. [1] introduced the problem of mining a large collection of basket data-type transactions for association rules between sets of items with minimal confidence threshold and presented an efficient algorithm for this purpose. Association rule mining has since been a research focus and has become a key technique in data mining, with broad and successful applications, such as in market analysis, financial investment, health, environmental protection, product manufacturing, and made a very significant social and economic benefit.

A key component in association rule mining problem (e.g., [2–4]) and other data mining problems such as episode [5] and minimal keys [6] is to discover frequent itemsets and maximal frequent itemsets. Maximal frequent itemset (MFI) can be used to improve the performance of discovering frequent itemsets. Furthermore, it suffices to know only the maximal frequent itemsets in many data mining applications, such as the minimal key discovery and the theory extraction [6].

* **Corresponding author: Jiang Liu,** Department of Systems Science, University of Shanghai for Science and Technology, Shanghai 200093, China, e-mail: jliu113@126.com
Jing Li, Feng Ni, Xiang Xia, Shunlong Li, Wenhui Dong: Department of Systems Science, University of Shanghai for Science and Technology, Shanghai 200093, China

Most of the research work with respect to discovering maximal frequent itemsets focused on developing deterministic algorithms (e.g., Bayardo [7]; Eppstein [8]; Lin and Kedem [9]; Boros et al. [10]; Dhabu and Deshpande [11]; Kabir et al. [12]; Karim et al. [13]; Halim et al. [14]), or complexity analysis (Eppstein [8]; Boros et al. [10]), or developing approximation algorithms (e.g., Fatemi et al. [15]; Zhang et al. [16]), or developing maximal frequent itemsets with added constraints such as maximal diverse frequent itemsets (e.g., Wu et al. [17]). Boros et al. [10] claimed that maximal t-frequent sets cannot be generated efficiently, unless P = NP. Eppstein [8] turned the problem of discovering maximal frequent itemsets into finding all maximal complete bipartite graphs of a bipartite graph on $m + n$ vertices and gave an algorithm with the complexity $O(l^3 4^l(m + n))$ in the worst case, where m, n are the item number and the transaction number, respectively, and l denotes the maximum of $|C||\Psi(C)|/(|C| + |\Psi(C)| - 1)$, with the maximum taken over all maximal frequent itemsets C.

In this article, we put forward a semigroup algebraic method for mining maximal frequent itemsets, whose complexity is $O(3mn2^\beta + 4^\beta n)$ (in practical applications), lower than the known complexity $O(l^3 4^l(m + n))$ given by Eppstein [8] both in the worst case, where β is the number of items whose support is more than the minimum support threshold.

First, we use algebraic language to describe the natural algebraic structure of items and transactions and explore some algebraic properties. For instance, there is a semigroup homomorphism between itemsets and transactions.

Under the algebraic framework, we give the explicit forms of simplified generators of frequent itemsets and prove that the simplified generators are maximal frequent itemsets and vice versa. Then we define basic itemsets and sim-basic itemsets and use them to construct the recurrence formula of maximal frequent itemsets. Finally, the recurrence formula is used to provide methods for discovering maximal frequent itemsets whose complexity proves to be much lower than the known complexity both in the worst case. Experimental comparisons also indicate a better performance of an algorithm based on the method.

2 Preliminaries

In this section, we briefly review some basic notations about algebraic semigroup and frequent itemset mining. Interested readers can refer to relative literature (e.g., Agrawal et al. [1]; Clifford and Preston [18]) for more details or more information with respect to the approaches to frequent itemset mining (e.g., Luna [19]).

Let $\mathcal{I}$ be a set of items, and $\mathcal{T}$ be a set of transactions, where each transaction has a unique identifier and contains a set of items (also called itemset). The support of an itemset X, denoted by $\text{supp}(X)$, is the number of transactions in which it occurs as a subset. An itemset is frequent if its support is not less than a pre-specified minimum support threshold $\min_{\text{sup}}$. X is a maximal frequent itemset if there is no other itemset X' such that X' is frequent and $X \subseteq X'$. X is a closed itemset if there is no other itemset X' such that $X \subset X'$ and $\text{supp}(X) = \text{supp}(X')$.

A *binary operation* on a set S is a mapping of $S \times S$ into S, where $S \times S$ is the set of all ordered pairs of elements of S. If the mapping is denoted by a dot $(\cdot)$, the image in S of the element (a, b) of $S \times S$ will be denoted by $a \cdot b$. Frequently, we shall omit the dot, writing ab for $a \cdot b$. Other symbols which may be used to denote binary operations are $+, \circ, *$.

A *semigroup* is a set S together with a binary operation $(\cdot)$, such that the operation is associative, i.e., $\forall a, b, c \in S, a \cdot (b \cdot c) = (a \cdot b) \cdot c$. Denote the semigroup by $S(\cdot)$ or S for simplicity when there is no danger of ambiguity, and say that S is a semigroup with respect to $\cdot$. If $a \cdot b = b \cdot a$, then S is called a commutative semigroup. If $e \in S$ and $\forall a \in S, a \cdot e = a = e \cdot a$, then e is an identity element.

A subset T of S is called a *sub-semigroup* of S if $a \in T$ and $b \in T$ imply $ab \in T$.

Suppose T is a non-empty subset of S. The sub-semigroup $\langle T \rangle$ is the set of all elements of S expressible as finite products of elements of T. T is called *a set of generators* of $\langle T \rangle$.

Let S and S' be groupoids. A mapping ϕ of S into S' is called a *homomorphism* if $(ab)\phi = (a\phi)(b\phi)$ for all a, b in S.

3 Algebraic properties

In this section, we investigate the algebraic semigroup structure and some basic algebraic properties for items and transactions.

In what follows, we always assume that **the element number of item set $\mathcal{I}$ is m, and the element number of transaction set $\mathcal{T}$ is n.**

We use $2^{\mathcal{I}}$ to denote the power set of $\mathcal{I}$, i.e., the set of all subsets of $\mathcal{I}$. And we denote the power set of $\mathcal{T}$ by $2^{\mathcal{T}}$.

For any two elements $X_1, X_2 \in 2^{\mathcal{I}}$, since $X_1 \bigcup X_2 \in 2^{\mathcal{I}}$, we can define the binary operation $\circ$ as follows:

$$X_1 \circ X_2 = X_1 \bigcup X_2.$$

$\circ$ has the following properties, for any $X_1, X_2, X_3 \in 2^{\mathcal{I}}$:

(i) $X_1 \circ (X_2 \circ X_3) = (X_1 \circ X_2) \circ X_3$ (associativity);
(ii) $\varnothing \circ X_1 = X_1$ (an identity element);
(iii) $X_1 \circ X_2 = X_2 \circ X_1$ (commutativity).

Hence, $2^{\mathcal{I}}$ is a commutative monoid with respect to the operation $\circ$, which has an identity element $\varnothing$.

Similarly, if we define the binary operation $*$ as $X_1 * X_2 = X_1 \bigcap X_2$, then $2^{\mathcal{I}}(*)$ is also a commutative monoid, which has an identity element $\mathcal{I}$.

Proposition 1. (Basic properties of itemsets) *Under $\circ$ and $*$, $2^{\mathcal{I}}$ has the following properties*, $\forall\, X, X_1, X_2, \ldots, X_k \in 2^{\mathcal{I}}$,

(*i*) $X_1 \circ (X_2 * X) = (X_1 \circ X_2) * (X_1 \circ X)$.
(*ii*) $X_1 * (X_2 \circ X) = (X_1 * X_2) \circ (X_1 * X)$.
(*iii*) $X \circ X = X, X * X = X$.
(*iv*) $(X_1 \circ X_2) * X_i = X_i$, $i = 1, 2$.
(*v*) *If* $X_1 * \ldots * X_k * X = X$, *then* $X_i * X = X$, $i = 1, 2, \ldots, k$.

From (i) and (iv), we directly have

$$(X_1 * X_2) \circ X_i = X_i, \quad i = 1, 2.$$

Similarly, $2^{\mathcal{T}}(\circ)$ and $2^{\mathcal{T}}(*)$ are also commutative monoids, which take $\varnothing$ and $\mathcal{T}$ as their identity elements.

The following mapping describes the support of an itemset X.

Definition 1. We define a mapping $\Psi : 2^{\mathcal{I}} \longrightarrow 2^{\mathcal{T}}$ as follows: For $X \in 2^{\mathcal{I}}$ and $X \neq \varnothing$, first find all the transactions $t_1, t_2, \ldots, t_s \in \mathcal{T}$, such that $X \subseteq t_i$. Then define $\Psi(X) = \{t_1, t_2, \ldots, t_s\}$. For $\varnothing$, define $\Psi(\varnothing) = \mathcal{T}$.

In fact, the mapping described in Definition 1 is a homomorphism.

Theorem 1. (Homomorphism between itemsets and transactions) Ψ *is a homomorphism, i.e., if* $X_1, X_2 \in 2^{\mathcal{I}}$, *then* $\Psi(X_1 \circ X_2) = \Psi(X_1) * \Psi(X_2)$.

Proof. If $\Psi(X_1) * \Psi(X_2) = \varnothing$, then obviously $\Psi(X_1 \circ X_2) \supseteq \Psi(X_1) * \Psi(X_2)$. Otherwise, let $t \in \Psi(X_1) * \Psi(X_2)$. Then $X_1, X_2 \subseteq t$. Hence, $X_1 \circ X_2 \subseteq t$, $t \in \Psi(a_1 \circ a_2)$, implying $\Psi(a_1 \circ a_2) \supseteq \Psi(a_1) * \Psi(a_2)$.

Conversely, if $\Psi(X_1 \circ X_2) = \varnothing$, then obviously $\Psi(X_1 \circ X_2) \subseteq \Psi(X_1) * \Psi(X_2)$. Otherwise, let $t \in \Psi(X_1 \circ X_2)$. Then $X_1 \circ X_2 \subseteq t$. Hence, $X_1, X_2 \subseteq t$, $t \in \Psi(X_1) * \Psi(X_2)$, implying $\Psi(X_1 \circ X_2) \subseteq \Psi(X_1) * \Psi(X_2)$. □

4 Simplified generators of frequent itemsets

In this section, we first present explicit forms of simplified generators for frequent itemsets. Then the maximal frequent itemsets prove to be the simplified generators.

Definition 2. Let $\min_{\sup}$ and **FI** be the minimal support threshold and the set of frequent itemsets, respectively. Suppose $\mathcal{U} \subseteq$ **FI**. For each element X in **FI**, there exists an element $\{I_1, \ldots, I_h\}$ in $\mathcal{U}$, such that $X \in \langle\{I_1\}, \ldots, \{I_h\}\rangle$. We say that $\mathcal{U}$ can generate all the frequent itemsets.

Lemma 1.
(1) *If* $|Y| \geq k$, *and* $Y \in \mathcal{Y}$, *then there exists* $Y_1 \in \mathcal{Y}$ *such that* $|Y_1| = k$, *and* $Y_1 * Y = Y_1$.
(2) *If* $Y_1 * Y_2 = Y_1$, *then* $|Y_1| \leq |Y_2|$.

Proof. (1) We can obtain a Y's subset which has k elements by removing $|Y| - k$ elements from Y. This completes the proof.

(2) It is obviously true by noting that Y_1 is a subset of Y_2. □

Theorem 2. (Simplified generators of frequent itemsets) *Let* $\min_{\sup}$ *be the minimal support threshold. First, find the elements* $Y_1, \ldots, Y_s$ *in* $\mathcal{Y}$ *such that* $|Y_i| = \min_{\sup}$, *and then find the elements* $I_1^i, \ldots, I_{h_i}^i$ *in* $\mathcal{I}$ *such that* $\Psi(\{I_j^i\}) * Y_i = Y_i$ *for* $1 \leq j \leq h_i$. *Let* **FI** *be the set of frequent itemsets.*

Furthermore, in the set $\Gamma = \{1, 2, \ldots, s\}$, *for any two integers* i, j, *if* $\{I_1^i, \ldots, I_{h_i}^i\}$ *is a subset of* $\{I_1^j, \ldots, I_{h_j}^j\}$, *then remove* i *from* Γ. *Finally, we obtain the set* Γ'. *Then*

(1) **FI** $= \bigcup_{i=1}^{s} \langle\{I_1^i\}, \ldots, \{I_{h_i}^i\}\rangle$. *In other words,* $\{\{I_1^i, \ldots, I_{h_i}^i\} | i = 1, \ldots, s\}$ *can generate all the frequent itemsets.* $\{I_1^i, \ldots, I_{h_i}^i\}$ *is called the G-frequent itemset for* Y_i.

(2) **FI** $= \bigcup_{i \in \Gamma'} \langle\{I_1^i\}, \ldots, \{I_{h_i}^i\}\rangle$. *We call* $\{\{I_1^i, \ldots, I_{h_i}^i\} | i \in \Gamma'\}$ *the set of simplified generators for* $\min_{\sup}$.

Proof. (1) Suppose $X_0 = \{I_{a1}, \ldots, I_{ap}\} \in$ **FI**, $I_{ai} \in \{I_1^k, \ldots, I_{h_k}^k\}$. Since $Y_k * \Psi(\{I_j^k\}) = Y_k$, $\Psi(X_0) * Y_k = Y_k$. Furthermore, $|\Psi(X_0)| \geq |Y_k| = \min_{\sup}$, implying that X_0 is a frequent itemset.

Conversely, suppose X_0 is a frequent itemset, where $X_0 = \{I_{a1}, \ldots, I_{ap}\}$, $I_{ai} \in \mathcal{I}$. Then $|\Psi(\{I_{a1}\}) * \Psi(\{I_{a2}\}) * \ldots * \Psi(\{I_{ap}\})| \geq \min_{\sup}$. By Lemma 1, there exists $Y_i \in \mathcal{Y}$ such that $|Y_i| = \min_{\sup}$, and $\Psi(\{I_{a1}\}) * \Psi(\{I_{a2}\}) * \ldots * \Psi(\{I_{ap}\}) * Y_i = Y_i$. Therefore, by Proposition 1, $\Psi(\{I_{aj}\}) * Y_i = Y_i$, which completes the proof.

(2) The statement is obviously true since for two subsets A, B of $\mathcal{X}$, if $A \subseteq B$, then $\langle A\rangle \subseteq \langle B\rangle$. □

Lemma 2. *Suppose* $\mathcal{U}$ *can generate all the frequent itemsets, and* $A \not\subseteq B$ *for any two elements* $A, B \in \mathcal{U}$. *Then* $\mathcal{U}$ *is the set of maximal frequent itemsets.*

Proof. This can be directly verified by the definition of maximal frequent itemset. □

The following statement shows that the maximal frequent itemsets are simplified generators.

Corollary 1. (Explicit forms of maximal frequent itemsets) *Let* $\{\{I_1^i, \ldots, I_{h_i}^i\} | i \in \Gamma'\}$ *be the set of simplified generators. Then it is the set of maximal frequent itemsets.*

Proof. By Theorem 2, $\{\{I_1^i, \ldots, I_{h_i}^i\} | i \in \Gamma'\}$ can generate all the frequent itemsets, and for any $j \neq i$, $\{I_1^i, \ldots, I_{h_i}^i\}$ is not a subset of $\{I_1^j, \ldots, I_{h_j}^j\}$. Hence, $\{\{I_1^i, \ldots, I_{h_i}^i\} | i \in \Gamma'\}$ is the set of maximal frequent itemsets by Lemma 2. □

5 Maximal frequent itemset discovering and complexity analysis

In this section, we first define basic itemset and sim-basic itemset and give a necessary and sufficient condition for a maximal $i+1$-frequent itemset being a subset of a closed i-frequent itemset, then provide a recurrence formula for maximal frequent itemsets. Finally, an algorithm for discovering maximal frequent itemsets is presented, based on the recurrence formula, and the complexity proves to be smaller than the known complexity given in [8] both in the worst case.

Given a minimal support threshold t, a maximal frequent itemset whose support is t is either a basic itemset or a sim-basic itemset, which will be proved in Theorem 3. Basic itemsets and sim-basic itemsets are defined as follows.

Definition 3. For $0 \le i \le n$, $\mathcal{I}^{(i)}$ denotes a subset of $\mathcal{I}$ such that for $I \in \mathcal{I}^{(i)}$, $|\Psi(I)| = i$. Divide $\mathcal{I}^{(i)}$ into disjoint subsets denoted by $X_1^{(i)}, \ldots, X_{i_s}^{(i)}$, such that for any two elements $I_1, I_2 \in \mathcal{I}^{(i)}$, $\Psi(\{I_1\}) = \Psi(\{I_2\})$ if and only if I_1, I_2 are in the same subset. For each subset $X_j^{(i)}$ $(1 \le j \le i_s)$, we use $\widetilde{X}_j^{(i)}$ to denote the itemset $\{I|\Psi(X_j^{(i)}) \subseteq \Psi(\{I\}), I \in \mathcal{I}\}$, and call it a *basic itemset* (*for* i).

Let $\mathcal{P}$ be the sub-semigroup of $(\mathcal{X}, \circ)$, generated by $\{\widetilde{X}_j^{(k)} | 1 \le j \le k_s, i+1 \le k \le n\}$. Suppose $X_0 \in \mathcal{P}$. If $|\Psi(X_0)| = i$, and for any basic itemset $\widetilde{X} \not\subseteq X_0$, $\Psi(X_0) \not\subseteq \Psi(\widetilde{X})$, then we call X_0 a *sim-basic itemset* (*for* i).

Remark 1. It is clear that $\Psi(\widetilde{X}_j^{(i)}) = \Psi(X_j^{(i)}) = \Psi(\{I_0\})$ by the definition of basic itemset, where $I_0 \in X_j^{(i)}$.

There is an equivalent description of basic itemsets and sim-basic itemsets as follows.

Proposition 2.
(1) *X is a basic itemset if and only if X is a closed itemset and there exists $I_0 \in \mathcal{I}$ such that $\Psi(\{I_0\}) = \Psi(X)$.*
(2) *X is a sim-basic itemset if and only if X is a closed itemset and $\Psi(\{I\}) \ne \Psi(X)$ for any $I \in \mathcal{I}$.*

Proof. (1) Suppose X is a basic itemset. By Definition 3, there exists $I_0 \in \mathcal{I}$ such that $\Psi(\{I_0\}) = \Psi(X)$. For $I \notin X$, $\Psi(X) \not\subseteq \Psi(\{I\})$ according to Definition 3. Hence, $\Psi(\{I\} \circ X) \ne \Psi(X)$, implying that X is a closed itemset.

Conversely, suppose X is a closed itemset and there exists $I_0 \in \mathcal{I}$ such that $\Psi(\{I_0\}) = \Psi(X)$. According to the definition of closed itemsets, $I_0 \in X$. Let X' be the basic itemset $\{I|\Psi(I_0) \subseteq \Psi(I)\}$. Then X' is a closed itemset. Since X and X' are both closed itemsets and $\Psi(X) = \Psi(X') = \Psi(\{I_0\})$, X is identical to X', implying that X is a basic itemset.

(2) Suppose X is a sim-basic itemset whose support is i $(1 \le i \le n)$. Assume that X is not a closed itemset, then there exists an itemset Y_0 such that $X \subset Y_0$ and $\Psi(Y_0) = \Psi(X)$. Hence, there exists an item I_0 such that $I_0 \in Y_0 - X$. Let $\widetilde{X}_0$ be the basic itemset $\{I|\Psi(\{I_0\}) \subseteq \Psi(\{I\}), I \in \mathcal{I}\}$. It is clear that $\widetilde{X}_0 \not\subseteq X$ since $I_0 \notin X$, but $\Psi(X) = \Psi(Y_0) \subseteq \Psi(\{I_0\}) = \Psi(\widetilde{X}_0)$, contradicting with Definition 3. Therefore, X is a closed itemset. Now we assume that there exists $I' \in \mathcal{I}$, such that $\Psi(\{I'\}) = \Psi(X)$. Then $I' \in X$, since X is a closed itemset. Therefore, there exists a basic itemset $\widetilde{X}_{j'}^{(k')}$ with $1 \le j' \le k'_s$, $i+1 \le k' \le n$, such that $I' \in \widetilde{X}_{j'}^{(k')}$, noting that X is in the sub-semigroup generated by $\{\widetilde{X}_j^{(k)} | 1 \le j \le i_s, i+1 \le k \le n\}$ by the definition of sim-basic itemset. This contradicts with the fact $|\Psi(\{I'\})| = |\Psi(X)| = i$.

Conversely, suppose $X = \{I_1, I_2, \ldots, I_q\}$ is a closed itemset whose support is i and $\Psi(\{I\}) \ne \Psi(X)$ for any $I \in \mathcal{I}$. Then $|\Psi(\{I_u\})| > |\Psi(X)| = i$ for $1 \le u \le q$. Let X' be $\widetilde{X}_1 \circ \widetilde{X}_2 \circ \ldots \circ \widetilde{X}_q$, where $\widetilde{X}_u$ is the basic itemset $\{I|\Psi(\{I_u\}) \subseteq \Psi(\{I\}), I \in \mathcal{I}\}$. Then $\Psi(X') = \Psi(X)$ by Remark 1. Note that X is a closed itemset. Hence, $X' = X$. Besides, $X' \in \mathcal{P}$ since $|\Psi(\widetilde{X}_u)| = |\Psi(\{I_u\})| > i$, where $\mathcal{P}$ is the sub-semigroup of $(\mathcal{X}, \circ)$, generated by $\{\widetilde{X}_j^{(k)} | 1 \le j \le k_s, i+1 \le k \le n\}$. Therefore, to prove that X is a sim-basic itemset, it suffices to prove that for any basic itemset $\widetilde{X} \not\subseteq X'$, $\Psi(X') \not\subseteq \Psi(\widetilde{X})$. Assume that there exists $\widetilde{X}_0 \not\subseteq X'$, such that $\Psi(X') \subseteq \Psi(\widetilde{X}_0)$. Since there exists $I_0 \in \widetilde{X}_0$ such that $\Psi(\{I_0\}) = \Psi(\widetilde{X}_0)$ by the definition of basic itemset, $I_0 \notin X'$ and $\Psi(X') \subseteq \Psi(\{I_0\})$. Hence, $\Psi(X' \circ \{I_0\}) = \Psi(X')$, contradicting with the fact that $X' = X$ is a closed itemset. □

The following is a preparation for Theorem 3.

Lemma 3. *Let $\mathcal{S}_i$, $\mathcal{B}_i$, and MFI_{i+1} be the set of sim-basic itemsets for i, the set of basic itemsets for i, and the set of maximal frequent itemsets when the minimum support threshold is $i+1$, respectively. Then each element in $\mathcal{B}_i$ cannot be a subset of any element in $\mathcal{S}_i \bigcup \mathrm{MFI}_{i+1}$. Each element in $\mathcal{S}_i$ cannot be a subset of any element in $\mathcal{B}_i \bigcup \mathrm{MFI}_{i+1}$.*

Proof. By Definition 3, for $X_0 \in \mathcal{B}_i$, there exists $I_0 \in \mathcal{I}$ such that $|\Psi(\{I_0\})| = i$ and $I_0 \in X_0$. And for $X_1 \in \mathcal{S}_i \bigcup \mathrm{MFI}_{i+1}$, $I_0 \notin X_1$ since X_1 either is in the sub-semigroup generated by $\{\widetilde{X}_j^{(k)} | 1 \le j \le k_s, i+1 \le k \le n\}$ by the definition of sim-basic itemset or $|\Psi(X_1)| > i$. Hence, X_0 cannot be a subset of X_1.

Suppose $X_0 \in \mathcal{B}_t$, $X_1 \in \mathcal{S}_i$, and $X_2 \in \mathrm{MFI}_{i+1}$. Then X_1 cannot be a subset of X_2 since $|\Psi(X_2)| \ge i+1$ and $|\Psi(X_1)| = i$. Since $X_0 \notin X_1$ as we have proved, $\Psi(X_1) \notin \Psi(X_0)$ by the definition of sim-basic itemset. Hence, $\Psi(X_1) \ne \Psi(X_0)$. Note that $|\Psi(X_1)| = |\Psi(X_0)| = i$. Therefore, X_1 cannot be a subset of X_0. □

Based on basic itemsets and sim-basic itemsets, the recurrence formula of maximal frequent itemsets can be deduced as follows.

Theorem 3. *Let MFI_{i+1}, MFI_i be the sets of maximal frequent itemsets when the minimum support thresholds are $i+1$ and i, respectively, then $\mathrm{MFI}_i = \widetilde{\mathrm{MFI}}_{i+1} \bigcup \mathcal{B}_i \bigcup \mathcal{S}_i$, where $\mathcal{S}_i$ and $\mathcal{B}_i$ have been defined in* Lemma 3, *and $\widetilde{\mathrm{MFI}}_{i+1}$ can be obtained by deleting the elements of MFI_{i+1} that are subsets of the elements in $\mathcal{S}_i \bigcup \mathcal{B}_i$.*

Proof. Suppose $X \in \mathrm{MFI}_i$. There are three cases.

Case 1. If $|\Psi(X)| = i$ and $\Psi(\{I\}) \ne \Psi(X)$ for any $I \in \mathcal{I}$, then X is a closed itemset since it is a maximal frequent itemset. Furthermore, X is a sim-basic itemset by Proposition 2, i.e., $X \in \mathcal{S}_i$.

Case 2. If $|\Psi(X)| = i$ and there exists $I_0 \in \mathcal{I}$ such that $\Psi(\{I_0\}) = \Psi(X)$, then X is the G-frequent itemset for $\Psi(X)$ by Theorem 2 and Corollary 1, i.e., $X = \{I | \Psi(X) \subseteq \Psi(\{I\}), I \in \mathcal{I}\}$. Let $\widetilde{X}_0$ be the basic itemset $\{I | \Psi(\{I_0\}) \subseteq \Psi(\{I\}), I \in \mathcal{I}\}$, which is exactly X. Hence, $X \in \mathcal{B}_i$.

Case 3. If $|\Psi(X)| > i$, then $X \in \widetilde{\mathrm{MFI}}_{i+1}$ for the following reasons.

Assume that $X \notin \widetilde{\mathrm{MFI}}_{i+1}$, then there exists a maximal frequent itemset for $\min_{\sup} = i+1$, denoted by X_0, such that $X \subset X_0$. Since X_0 is also a frequent itemset for $\min_{\sup} = i$, X_0 is a subset of a maximal frequent itemset for $\min_{\sup} = i$. Therefore, X is a proper subset of a maximal frequent itemset for $\min_{\sup} = i$, which contradicts with the fact that X is a maximal frequent itemset for $\min_{\sup} = i$.

Now we claim that $\widetilde{\mathrm{MFI}}_{i+1} \bigcup \mathcal{B}_i \bigcup \mathcal{S}_i$ generate all the frequent itemsets. This can be verified by the fact that each element in $\widetilde{\mathrm{MFI}}_{i+1} \bigcup \mathcal{B}_i \bigcup \mathcal{S}_i$ is a frequent itemset for $\min_{\sup} = i$, and $\mathrm{MFI}_i \subseteq \widetilde{\mathrm{MFI}}_{i+1} \bigcup \mathcal{B}_i \bigcup \mathcal{S}_i$ as we have proved. Besides, $X_1 \not\subseteq X_2$ for $X_1, X_2 \in \widetilde{\mathrm{MFI}}_{i+1} \bigcup \mathcal{B}_i \bigcup \mathcal{S}_i$ by Lemma 3. Therefore, $\widetilde{\mathrm{MFI}}_{i+1} \bigcup \mathcal{B}_i \bigcup \mathcal{S}_i$ is the set of maximal frequent itemsets for $\min_{\sup} = i$ by Lemma 2. □

Lemma 4. *Suppose $X \in \mathrm{MFI}_{i+1}$. Then X is a subset of an element in $\mathcal{S}_i \bigcup \mathcal{B}_i$, if and only if there exists $I_0 \in \mathcal{I}$ such that $|\Psi(X \circ \{I_0\})| = i$.*

Proof. Suppose there exists $I_0 \in \mathcal{I}$ such that $|\Psi(X \circ \{I_0\})| = i$. Then X must be a subset of a closed itemset whose support is i. Since $\mathcal{S}_i \bigcup \mathcal{B}_i$ is the set of closed itemsets whose support is i according to Proposition 2, X is a subset of an element in $\mathcal{S}_i \bigcup \mathcal{B}_i$.

Suppose X is a subset of an element in $\mathcal{S}_i \bigcup \mathcal{B}_i$. Then there exists an itemset X_0 such that $|\Psi(X \circ X_0)| = i$ and $X \circ X_0$ is a closed itemset. Let I_0 be an element in X_0. Assume that $|\Psi(X \circ \{I_0\})| < i$, then $|\Psi(X_0 \circ X)| \le |\Psi(X \circ \{I_0\})| < i$, contradicting with that $|\Psi(X \circ X_0)| = i$. Hence, $|\Psi(X \circ \{I_0\})| \ge i$. Since $X \in \mathrm{MFI}_{i+1}$, $|\Psi(X \circ \{I_0\})| < i+1$ by the definition of maximal frequent itemsets. Finally, we obtain that $|\Psi(X \circ \{I_0\})| = i$, completing the proof. □

Definition 4. Let $\mathcal{B}_i$ and $\mathcal{S}_i$ be the set of basic itemsets and sim-basic itemsets for i, respectively. Find the subset of $\mathcal{B}_i$ (or $\mathcal{S}_i$) such that for each element X in the subset, there exists $I \in \mathcal{I}$ such that $|\Psi(\{I\} \circ X)| = k$ with $0 \le k \le i-1$, then the subset is called a Re-k subset of $\mathcal{B}_i$ (or $\mathcal{S}_i$), and denoted by $\mathcal{B}_i^k$ (or $\mathcal{S}_i^k$).

Theorem 3 and Lemma 4 immediately lead to the following theorem.

Theorem 4. *Given a minimum support threshold t, suppose $t+1 \le i \le n$, for $k = t, t+1, \ldots, i-1$, delete the Re-k subset from $\mathcal{B}_i$ (or $\mathcal{S}_i$), and obtain a new set denoted by $\widetilde{\mathcal{B}}_i$ (or $\widetilde{\mathcal{S}}_i$). Then* $\mathrm{MFI}_t = \bigcup_{i=t+1}^{n} (\widetilde{\mathcal{B}}_i \bigcup \widetilde{\mathcal{S}}_i) \bigcup \mathcal{B}_t \bigcup \mathcal{S}_t$.

Proof. $\mathrm{MFI}_n = \mathcal{B}_n$ and $\mathcal{S}_n = \varnothing$. According to Lemma 4 and Theorem 3, $\widetilde{\mathrm{MFI}}_n$ can be obtained by deleting the Re-$(n-1)$ subset of $\mathcal{B}_n$, and $\mathrm{MFI}_{n-1} = (\mathcal{B}_n - \mathcal{B}_n^{n-1}) \bigcup \mathcal{B}_{n-1} \bigcup \mathcal{S}_{n-1}$. Similarly, $\mathrm{MFI}_{n-2} = (\mathcal{B}_n - \mathcal{B}_n^{n-1} - \mathcal{B}_n^{n-2}) \bigcup (\mathcal{B}_{n-1} - \mathcal{B}_{n-1}^{n-2}) \bigcup (\mathcal{S}_{n-1} - \mathcal{S}_{n-1}^{n-2}) \bigcup \mathcal{B}_{n-2} \bigcup \mathcal{S}_{n-2}$. Generally,

$$\mathrm{MFI}_s = \bigcup_{j=s+1}^{n} (\mathcal{B}_j - \mathcal{B}_j^{j-1} - \ldots - \mathcal{B}_j^{s}) \bigcup (\mathcal{S}_j - \mathcal{S}_j^{j-1} - \ldots - \mathcal{S}_j^{s}) \bigcup \mathcal{B}_s \bigcup \mathcal{S}_s,$$

where $0 \le s \le n-1$, completing the proof. □

Proposition 3. *Let i_s $(0 \le i \le n)$ be the number of disjoint subsets of the set $\mathcal{I}_i = \{I | I \in \mathcal{I}, |\Psi(I)| = i\}$ as defined in* Definition 3.

(1) $0_s + 1_s + 2_s + \ldots + n_s \le m$.

(2) $\sum_{i=p}^{n} |\mathcal{S}_i| \le |\mathcal{P}| \le 2^{(p+1)_s + (p+2)_s + \ldots + n_s} - 1$, *where $\mathcal{P}$ is the sub-semigroup of $(\mathcal{X}, \circ)$, generated by $\{\widetilde{X}_j^{(k)} | 1 \le j \le k_s, p+1 \le k \le n\}$.*

Proof. (1) Straightforward by the definitions of basic itemsets.

(2) By Definition 3, $\sum_{i=p}^{n} |\mathcal{S}_i| \le |\mathcal{P}|$, and $|\mathcal{P}|$ is

$$C_{(p+1)_s + (p+2)_s + \ldots + n_s}^{1} + C_{(p+1)_s + \ldots + n_s}^{2} + \ldots + C_{(p+1)_s + \ldots + n_s}^{(p+1)_s + \ldots + n_s} = 2^{(p+1)_s + \ldots + n_s} - 1,$$

since for any basic itemset $\widetilde{X}_j^{(k)}$, the product of arbitrary number of $\widetilde{X}_j^{(k)}$'s is still $\widetilde{X}_j^{(k)}$. □

By Theorem 3, we directly have the following algorithm for discovering maximal frequent itemsets.

Algorithm 1. Discovery of maximal frequent itemsets.

Input: $\mathrm{MFI}_n = \{I | \Psi(\{I\}) = \mathcal{T}, I \in \mathcal{I}\}$, and the minimum support threshold t.

Output: MFI_t.

Step 1. For $t \le i \le n$, find the items from $\mathcal{I}$ such that the support of each item is i. Denote the set of these items by $\mathcal{I}^{(i)}$.

Step 2. Divide $\mathcal{I}^{(i)}$ into disjoint subsets $X_1^{(i)}, \ldots, X_{i_s}^{(i)}$ such that for any two elements $I_1, I_2 \in \mathcal{I}^{(i)}$, $\Psi(\{I_1\}) = \Psi(\{I_2\})$ if and only if I_1, I_2 are in the same subset.

Step 3. For $X_j^{(i)}$ $(1 \le j \le i_s, t \le i \le n)$, find the subset $\{I | \Psi(X_j^{(i)}) \subseteq \Psi(\{I\}), I \in \mathcal{I}\}$. Denote it by $\widetilde{X}_j^{(i)}$.

Step 4. Discovery of sim-basic itemsets.

4.1 Let $\widetilde{\mathbb{X}}$ be the set $\{\widetilde{X}_j^{(i)} | i = t+1, \ldots, n, j = 1, 2, \ldots, i_s\}$, and $i_0 = t+1$.

4.2 For $\widetilde{X}_j^{(i_0)} \in \widetilde{\mathbb{X}}$, find the subset $\{\widetilde{X}_j^{(k)} | \Psi(\widetilde{X}_j^{(i_0)}) \subseteq \Psi(\widetilde{X}_j^{(k)}), \widetilde{X}_j^{(k)} \in \widetilde{\mathbb{X}}\}$, denoted by $H_j^{(i_0)}$.

4.3 Let $\widetilde{\mathbb{X}} = \widetilde{\mathbb{X}} - \bigcup_{j=1}^{i_s} H_j^{(i_0)}$, and $i_0 = i_0 + 1$. If $\widetilde{\mathbb{X}} \ne \varnothing$, return to Step 4.2.

4.4 Let $P_j^{(i_0)}$ be the product of elements in $H_j^{(i_0)}$ under $\circ$, and $\mathcal{P}$ be the sub-semigroup of $(\mathcal{X}, \circ)$, generated by $\{P_j^{(i)} | 1 \le j \le i_s, t+1 \le i \le n\}$.

4.5 Denote $\mathcal{P} - \{P_j^{(i)} | 1 \le j \le i_s, t+1 \le i \le n\}$ by $\mathcal{P}'$. Divide $\mathcal{P}'$ into disjoint subsets $\{\mathcal{P}'_1, \mathcal{P}'_2, \ldots, \mathcal{P}'_u\}$ such that for $A, B \in \mathcal{P}'$, A and B are in the same subset if and only if $\Psi(A) = \Psi(B)$. For $\mathcal{P}'_i$, let X_i be the product of all the elements in $\mathcal{P}'_i$ under $\circ$. Then let $\mathcal{M}$ be $\{X_1, \ldots, X_u\}$.

Step 5. Computation of MFI_t.

5.1 For $t \le i \le n$, let $\mathcal{B}_i$ be $\{\widetilde{X}_j^{(i)} | 1 \le j \le i_s\}$, and let $\mathcal{S}_i$ be $\{X_j | |\Psi(X_j)| = i, X_j \in \mathcal{M}\}$.

5.2 For $t+1 \le i \le n$ and $X \in \mathcal{B}_i$ (or $\mathcal{S}_i$), if there exists $I \in \mathcal{I}$ such that $|\Psi(\{I\} \circ X)| = k$ with $t \le k \le i-1$, then delete X from $\mathcal{B}_i$ (or $\mathcal{S}_i$) to obtain a new set denoted by $\widetilde{\mathcal{B}}_i$ (or $\widetilde{\mathcal{S}}_i$).

5.3 Output $\mathrm{MFI}_t = \bigcup_{i=t+1}^{n} (\widetilde{\mathcal{B}}_i \bigcup \widetilde{\mathcal{S}}_i) \bigcup \mathcal{B}_t \bigcup \mathcal{S}_t$.

Proposition 4. *Let $P_{j1}^{(i1)} \circ P_{j2}^{(i1)} \circ \ldots \circ P_{jq}^{(iq)}$ be an element in $\mathcal{P}$ with $q \ge 2$. Then for any $I \in \mathcal{I}$, we have $\Psi(P_{j1}^{(i1)} \circ P_{j2}^{(i1)} \circ \ldots \circ P_{jq}^{(iq)}) \ne \Psi(I)$, where P_{j1}^{i1} is the product of elements in $H_{j1}^{(i1)}$ under $\circ$, and $\mathcal{P}$ is the semigroup generated by $\{P_j^{(i)} | 1 \le j \le i_s, t+1 \le i \le n\}$, as defined in Step 4 of Algorithm* 1.

Proof. Steps 4.1–4.3 imply that for $H_j^{(i)}$, there does not exist $I \in \mathcal{I}$ such that $\Psi(I) \subset \Psi(P_j^{(i)})$, where $P_j^{(i)}$ is the product of elements in $H_j^{(i)}$ under $\circ$. Therefore, for any two different sets $H_j^{(i)}$ and $H_{j0}^{(i0)}$, $\Psi(P_j^{(i)}) \not\subseteq \Psi(P_{j0}^{(i0)})$ and $\Psi(P_{j0}^{(i0)}) \not\subseteq \Psi(Pj^{(i)})$. Consequently, there does not exist $I \in \mathcal{I}$ such that $\Psi(I) = \Psi(P_{j1}^{(i1)} \circ P_{j2}^{(i1)} \circ \ldots \circ P_{jq}^{(iq)})$, completing the proof. □

Remark 2.

(1) Steps 1–3 are to find basic itemsets.

(2) According to Proposition 4, Step 4 is to find closed itemsets which meet the condition in Proposition 2, hence Step 4 can discover sim-basic itemsets under $\circ$.

Example 1. Suppose a database has six transactions and five items, as shown in Table 1. Let the minimum support threshold t be 3. Then MFI_3 can be obtained by Algorithm 1.

Step 1. For $3 \le i \le 6$, find the set of items whose support is i, and denote it by $\mathcal{I}^{(i)}$. We have $\mathcal{I}^{(4)} = \{B, C\}$, and $\mathcal{I}^{(5)} = \{D, E\}$.

Step 2. Divide $\mathcal{I}^{(4)}$ into disjoint subsets $X_1^{(4)} = \{B\}$, $X_2^{(4)} = \{C\}$. Similarly, $X_1^{(5)} = \mathcal{I}^{(5)} = \{D, E\}$.

Step 3. For $X_j^{(i)}$ $(1 \le j \le i_s, t \le i \le n)$, find the subset $\widetilde{X}_j^{(i)} = \{I | \Psi(X_j^{(i)}) \subseteq \Psi(\{I\}), I \in \mathcal{I}\}$. We have $\widetilde{X}_1^{(4)} = \{B, D, E\}$, $\widetilde{X}_2^{(4)} = \{C\}$, and $\widetilde{X}_1^{(5)} = \{D, E\}$.

Step 4. $H_1^{(4)} = \{\widetilde{X}_1^{(4)}, \widetilde{X}_1^{(5)}\}$, $H_2^{(4)} = \{\widetilde{X}_2^{(4)}, \widetilde{X}_1^{(5)}\}$, and $H_1^{(5)} = \{\widetilde{X}_1^{(5)}\}$. Then $P_1^{(4)} = \widetilde{X}_1^{(4)} \circ \widetilde{X}_1^{(5)}$, $P_2^{(4)} = \widetilde{X}_2^{(4)} \circ \widetilde{X}_1^{(5)}$, and $P_1^{(5)} = \{\widetilde{X}_1^{(5)}\}$.

Step 5. $\mathcal{P}' = \mathcal{M} = \{\widetilde{X}_1^{(4)} \circ \widetilde{X}_2^{(4)} \circ \widetilde{X}_1^{(5)}\}$.

Step 6. From Steps 4 and 5, $\mathcal{B}_3 = \varnothing$, $\mathcal{S}_3 = \{\widetilde{X}_1^{(4)} \circ \widetilde{X}_2^{(4)} \circ \widetilde{X}_1^{(5)}\}$, $\mathcal{B}_4 = \{\widetilde{X}_1^{(4)}, \widetilde{X}_2^{(4)}\}$, $\mathcal{S}_4 = \varnothing$, $\mathcal{B}_5 = \{\widetilde{X}_1^{(5)}\}$, $\mathcal{S}_5 = \varnothing$.

Step 7. Since $|\Psi(\widetilde{X}_1^{(4)} \circ \{C\})| = 3$ and $|\Psi(\widetilde{X}_2^{(4)} \circ \{B\})| = 3$, $\widetilde{\mathcal{B}}_4 = \varnothing$. And $\widetilde{\mathcal{B}}_5 = \varnothing$ since $|\Psi(\widetilde{X}_1^{(5)} \circ \{B\})| = 4$.

Step 8. Output $\mathrm{MFI}_3 = \mathcal{S}_3 = \{\widetilde{X}_1^{(4)} \circ \widetilde{X}_2^{(4)} \circ \widetilde{X}_1^{(5)}\} = \{\{B, C, D, E\}\}$, since $\mathcal{B}_3 = \varnothing$, and $\widetilde{\mathcal{B}}_i = \widetilde{\mathcal{S}}_i = \varnothing$ for $i = 4, 5$.

Lemma 5. *Suppose A, B are two sorted sets in ascending order, and $|A| = p$, $|B| = q$, then the complexity of computing $A \bigcap B$ is $p + q$.*

Table 1: Sample database

Transactions	Items
1	BCDE
2	ABDE
3	BCDE
4	ABCDE
5	DE
6	C

Proof. $A \bigcap B$ can be obtained by carrying out the following steps.

Step 1. Let $C = \varnothing$.

Step 2. If $A \neq \varnothing$ and $B \neq \varnothing$, carry out Step 3. Otherwise, output C.

Step 3. Suppose a_1 and b_1 are the first elements of A and B, respectively. If $a_1 < b_1$, then delete a_1 to obtain a new set denoted by A and return to Step 2. If $a_1 = b_1$, then add a_1 to C, delete a_1 and b_1 from A and B, respectively, to obtain new sets denoted by A and B, respectively, and return to Step 2. If $a_1 > b_1$, then delete b_1 from B to obtain a new set denoted by B, and return to Step 2.

Since at least one element is deleted from $A \bigcap B$ in each Step 3, the complexity of computing $A \bigcap B$ is $p + q$. □

Proposition 5. *The complexity of discovering maximal frequent itemsets for* $\min_{\sup} = t$ *is less than* $O(3mn2^{\beta} + 4^{\beta}n)$, *where* β *is* $(t+1)_s + \ldots + n_s$, *and* i_s $(t+1 \le i \le n)$ *is the number of disjoint subsets of the set* $\mathcal{I}_i = \{I | I \in \mathcal{I}, |\Psi(I)| = i\}$ *as defined in* Definition 3.

Proof. We define an order on $\mathcal{T}$ as $t_i < t_j$ if and only if $i < j$.

First of all, for $I \in \mathcal{I}$, $\Psi(\{I\})$ can be obtained as a sorted set in ascending order by comparing I with the elements in $t_i \in \mathcal{T}$ from $i = 1$ to $i = n$, hence the complexity of computing $\{\Psi(\{I\}) | I \in \mathcal{I}\}$ is m^2n.

Steps 1 and 2 of Algorithm 1 are to divide $\mathcal{I}$ into disjoint subsets by comparing $\Psi(\{I_1\})$ with $\Psi(\{I_2\})$ for every two elements I_1, I_2 in $\mathcal{I}$. So the complexity is at most $2nC_m^2$ by Lemma 5.

Step 3 is to find $\widetilde{X}_j^{(i)}$ also by making comparisons between $\Psi(X_j^{(i)})$ and $\Psi(\{I\})$. And we do not need to compute $\Psi(X_j^{(i)})$ since $\Psi(X_j^{(i)}) = \Psi(\{I\})$ for any element I in $X_j^{(i)}$. So the complexity is also at most $2nC_m^2$ by Lemma 5. Besides, there is no need to compute $\Psi(\widetilde{X}_j^{(i)})$ in the following Step 4, since $\Psi(\widetilde{X}_j^{(i)}) = \Psi(X_j^{(i)})$ by Remark 1.

In Step 4, Steps 4.1–4.3 make comparisons between every two elements in $\widetilde{\mathbb{X}}$, hence the complexity is at most C_{β}^2, where β denotes $(t+1)_s + \ldots + n_s$.

The complexity of computing $\{\Psi(A) | A \in \mathcal{P}'\}$ is at most $(2^{\beta} - 1)n\beta$, noting that the element number of $\mathcal{P}'$ is less than $2^{\beta} - 1$ by Proposition 3 and $\Psi(A)$ is an intersection of some elements in $\{\Psi(\widetilde{X}_j^{(i)}) | 1 \le j \le i_s, t+1 \le i \le n\}$ since $\Psi(P_j^i) = \Psi(\widetilde{X}_j^i)$.

The complexity of Step 4.5 is less than $2nC_{2^{\beta}-1}^2$, since Step 4.5 compares $\Psi(A)$ with $\Psi(B)$ for every two elements A, B in $\mathcal{P}'$.

Step 5 is to compute $\Psi(\{I\} \circ X) = \Psi(\{I\}) \bigcap \Psi(X)$, where $I \in \mathcal{I}$, and X is an element in $\mathcal{B}_i \bigcup \mathcal{S}_i$ with $t+1 \le i \le n$. Hence, the complexity of Step 5 is at most $\sum_{i=t+1}^{n}(|\mathcal{B}_i| + |\mathcal{S}_i|)m(n+n)$, which is less than $2(\beta + 2^{\beta})mn$ according to Proposition 3.

Therefore, the total complexity of Algorithm 1 is less than $m^2n + 4C_m^2n + C_{\beta}^2 + 2^{\beta}n\beta + 2C_{2^{\beta}-1}^2n + 2(\beta + 2^{\beta})mn$, or $O(3mn2^{\beta} + 4^{\beta}n)$. □

Remark 3. Eppstein [8] gave an $O(l^3 4^l(m+n))$ algorithm to generate all maximal frequent itemsets by graph theory, where l denotes the maximum of $|C||\Psi(C)|/(|C| + |\Psi(C)| - 1)$, with the maximum taken over all maximal frequent itemsets C. Since $n, t \gg m$ in practical applications, $|\Psi(C)| \gg |C|$. Hence, $|C||\Psi(C)|/(|C| + |\Psi(C)| - 1)$ is nearly m in the worst case. The complexity of Algorithm 1 in the worst case is $O(3mn2^m + 4^m n)$, which is much lower than $O(m^3 4^m(m+n))$.

6 Experiments

Besides the theoretical analysis, we use Python 3.5 to evaluate the performance of Algorithm 1 on four different datasets *Chess*, *Mushroom*, $T40I10D100K$, and $T10I4D100K$ from UCI that have been commonly used in previous research. All experiments are performed on a computer having Microsoft Windows 7

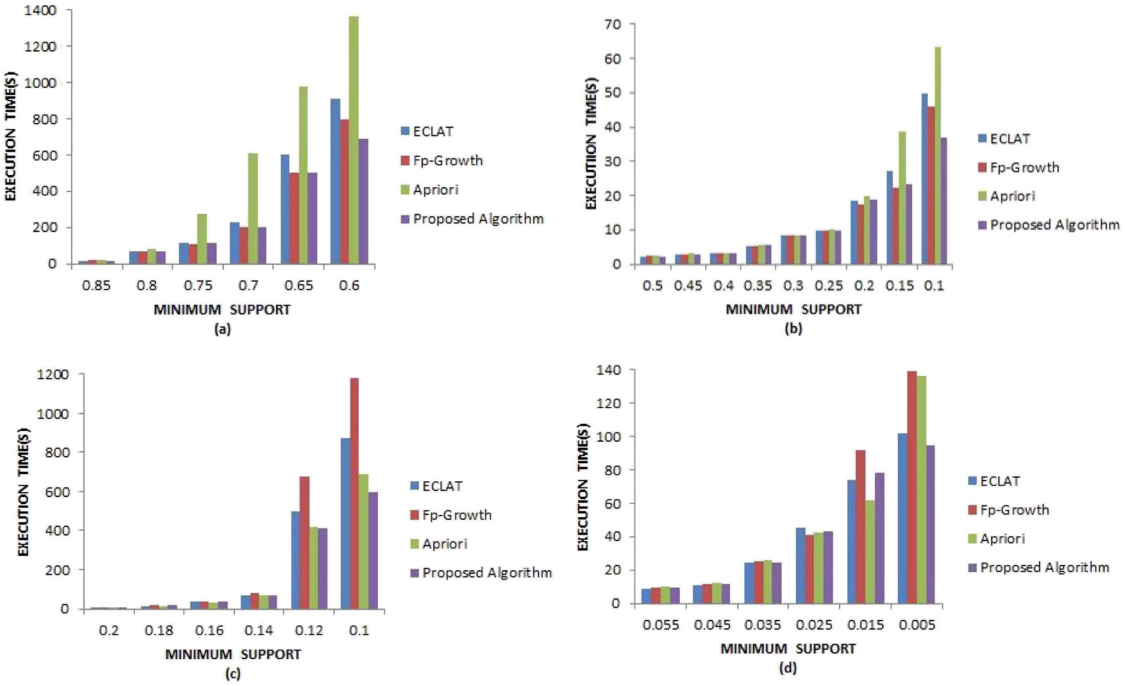

Figure 1: Execution time vs minimum support (a) using *Chess* dataset, (b) using *Mushroom* dataset, (c) using $T40I10D100K$ dataset, and (d) using $T10I4D100K$ dataset.

operating system, a core i7 CPU, and 8 GB of RAM. Our algorithm is compared with the three well-known deterministic algorithms for frequent itemset mining: ECLAT [20], Apriori [21], and FP-Growth [22].

Since we mainly considered complexities of algorithms in this article, we use the execution time as the performance measure. We examined each dataset with various minimum support thresholds and saved the execution time of each algorithm.

The results in Figure 1 indicate that Algorithm 1 is performing better than the other approaches considering their execution time.

7 Conclusion

We have presented methods to discover maximal frequent itemsets from the perspective of semigroup algebra and proved that the complexity of the methods is much lower than the known complexity in the worst case. Experiments made on four commonly used datasets also show that the algorithm based on our method performs better than the other three well-known algorithms. Meanwhile, we provided explicit forms of simplified generators of frequent itemsets, proved that the simplified generators are maximal frequent itemsets and vice versa, provided a necessary and sufficient condition for a maximal $i+1$-frequent itemset being a subset of a closed i-frequent itemset, and put forward a recurrence formula of maximal frequent itemsets by defining basic itemsets and sim-basic itemsets.

We also explored some algebraic properties of rule mining, which can be used to investigate other basic problems such as more efficient algorithms for discovering closed frequent itemsets, generators of association rules and reducing redundant association rules in further work.

Acknowledgments: The authors sincerely thank the referees for their constructive comments and fruitful suggestions that helped improving this paper.

Funding information: This work was partly supported by National Natural Science Foundation of China (Grant No. 11701370).

Author contributions: Jiang Liu and Feng Ni contributed in an overall design and the writing of this paper, put forward the main results and gave the proof; Jing Li, Xiang Xia, Shunlong Li, and Wenhui Dong contributed in the experiments.

Conflict of interest: The authors state no conflict of interest.

References

[1] R. Agrawal, T. Imieliński, and A. Swami, *Mining association rules between sets of items in large databases*, ACM SIGMOD Record **22** (1993), no. 2, 207–216, DOI: https://doi.org/10.1145/170036.170072.

[2] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A. I. Verkamo, *Fast Discovery of Association Rules: Advances in Knowledge Discovery and Data Mining*, MIT Press, California, 1996, pp. 307–328.

[3] J. Han and Y. Fu, *Discovery of multiple-level association rules from large databases*, in: VLDB '95 Proceedings of the 21th International Conference on Very Large Data Bases, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1995, pp. 420–431.

[4] W. Hwang and D. Kim, *Improved association rule mining by modified trimming*, in: The Sixth IEEE International Conference on Computer and Information Technology (CIT'06), IEEE Computer Society, Los Alamitos, CA, USA, 2006, pp. 24–24, DOI: https://doi.org/10.1109/CIT.2006.101.

[5] H. Mannila, H. Toivonen, and A. I. Verkamo, *Discovering frequent episodes in sequences*, in: Proceedings of First ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), AAAI Press, Palo Alto, CA, USA, 1995, pp. 210–215.

[6] D. Gunopulos, H. Mannila, and S. Saluja, *Discovering all most specific sentences by randomized algorithm*, in: F. Afrati, P. Kolaitis (eds), Database Theory - ICDT '97, Lecture Notes in Computer Science, Vol 1186. Springer, Berlin, Heidelberg, 1997.

[7] R. J. Bayardo, *Efficiently mining long patterns from databases*, ACM SIGMOD Record **27** (1998), no. 2, 85–93, DOI: https://doi.org/10.1145/276305.276313.

[8] D. Eppstein, *Arboricity and bipartite subgraph listing algorithms*, Inform. Process. Lett. **51** (1994), no. 4, 207–211.

[9] D. Lin and Z. M. Kedem, *Pincer-search: an efficient algorithm for discovering the maximum frequent set*, IEEE Trans. Knowl. Data Eng. **14** (2002), no. 3, 553–566, DOI: https://doi.org/10.1109/TKDE.2002.1000342.

[10] E. Boros, V. Gurvich, L. Khachiyan, and K. Makino, *On maximal frequent and minimal infrequent sets in binary matrices*, Ann. Math. Artif. Intell. **39** (2003), 211–221, DOI: https://doi.org/10.1023/A:1024605820527.

[11] M. M. Dhabu and P. S. Deshpande, *Cardinality statistics based maximal frequent itemsets mining*, in: S. Dua, A. Gangopadhyay, P. Thulasiraman, U. Straccia, M. Shepherd, B. Stein (eds), Information Systems, Technology and Management. Communications in Computer and Information Science, Vol. 285, Springer, Berlin, Heidelberg, 2021, DOI: https://doi.org/10.1007/978-3-642-29166-1_3.

[12] M. M. J. Kabir, S. Xu, B. H. Kang, and Z. Zhao, *Comparative analysis of genetic based approach and Apriori algorithm for mining maximal frequent item sets*, in: 2015 IEEE Congress on Evolutionary Computation (CEC), 2015, pp. 39–45, DOI: https://doi.org/10.1109/CEC.2015.7256872.

[13] M. R. Karim, M. Cochez, O. D. Beyan, C. F. Ahmed, and S. Decker, *Mining maximal frequent patterns in transactional databases and dynamic data streams: A spark-based approach*, Inf. Sci. **432** (2018), 278–300, DOI: https://doi.org/10.1016/j.ins.2017.11.064.

[14] Z. Halim, O. Ali, and M. G. Khan, *On the efficient representation of datasets as graphs to mine maximal frequent itemsets*, IEEE Trans. Knowl. Data Eng. **33** (2021), no. 4, 1674–1691, DOI: https://doi.org/10.1109/TKDE.2019.2945573.

[15] S. M. Fatemi, S. M. Hosseini, A. Kamandi, and M. Shabankhah, *CL-MAX: a clustering-based approximation algorithm for mining maximal frequent itemsets*, Int. J. Mach. Learn. Cybern. **12** (2021), no. 2, 365–383, DOI: https://doi.org/10.1007/s13042-020-01177-5.

[16] Y. Zhang, W. Yu, X. Ma, H. Ogura, and D. Ye, *Multi-objective optimization for high-dimensional maximal frequent itemset mining*, Appl. Sci. **11** (2021), no. 19, 8971, DOI: https://doi.org/10.3390/app11198971.

[17] D. Wu, D. Luo, C. S. Jensen, and J. Z. Huang, *Efficiently mining maximal diverse frequent itemsets*, in: G. Li, J. Yang, J. Gama, J. Natwichai, Y. Tong (eds), Database Systems for Advanced Applications. Lecture Notes in Computer Science, Vol 11447, Springer, Cham, 2019, DOI: https://doi.org/10.1007/978-3-030-18579-4_12.

[18] A. H. Clifford and G. B. Preston, *The Algebraic Theory of Semigroups*, American Mathematical Society, Providence, Rhode Island, 1961.
[19] J. M. Luna, P. Fournier-Viger, and S. Ventura, *Frequent itemset mining: A 25 years review*, Wiley Interdiscip. Rev. Data Min. Knowl. Discov. **9** (2019), no. 6, e1329, DOI: https://doi.org/10.1002/widm.1329.
[20] M. J. Zaki, *Scalable algorithms for association mining*, IEEE Trans. Knowl. Data. Eng. **12** (2000), no. 3, 372–390, DOI: https://doi.org/10.1109/69.846291.
[21] R. Agrawal and R. Srikant, *Fast algorithms for mining association rules*, in: Proceedings of the 20th International Conference on Very Large Data Bases (VLDB '94), Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1994, pp. 487–499.
[22] J. Han, J. Pei, and Y. Yin, *Mining frequent patterns without candidate generation*, SIGMOD Rec. **29** (2000), no. 2, 1–12, DOI: https://doi.org/10.1145/335191.335372.