Data Mining and Knowledge Discovery, 8, 227–252, 2004

Pushing Convertible Constraints in Frequent Itemset Mining*

JIAN PEI jianpei@cse.buffalo.edu
University at Buffalo, The State University of New York, 201 Bell Hall, Buffalo, NY 14260-2000, USA

JIAWEI HAN hanj@cs.uiuc.edu
University of Illinois at Urbana-Champaign, 2123 DCL, 1304 West Springfield Avenue, Urbana, IL 61801, USA

LAKS V.S. LAKSHMANAN laks@cs.ubc.ca
University of British Columbia, 201-2366 Main Mall, Vancouver, B.C. Canada V6T 1Z4

Editors: Fayyad, Mannila, Ramakrishnan

Received July 12, 2001; Revised July 12, 2001

Abstract. Recent work has highlighted the importance of the constraint-based mining paradigm in the context of frequent itemsets, associations, correlations, sequential patterns, and many other interesting patterns in large databases. Constraint pushing techniques have been developed for mining frequent patterns and associations with antimonotonic, monotonic, and succinct constraints. In this paper, we study constraints which cannot be handled with existing theory and techniques in frequent pattern mining. For example, $avg(S)\theta v$, $median(S)\theta v$, $sum(S)\theta v$ (S can contain items of arbitrary values, $\theta \in \{>, <, \leq, \geq\}$ and v is a real number.) are customarily regarded as "tough" constraints in that they cannot be pushed inside an algorithm such as Apriori. We develop a notion of *convertible constraints* and systematically analyze, classify, and characterize this class. We also develop techniques which enable them to be readily pushed deep inside the recently developed FP-growth algorithm for frequent itemset mining. Results from our detailed experiments show the effectiveness of the techniques developed.

Keywords: frequent itemset mining, constraint, convertible constraint, algorithm, pruning

1. Introduction

It has been well recognized that frequent pattern mining plays an essential role in discovering associations (Agrawal and Srikant, 1994; Mannila et al., 1994), correlations (Brin et al., 1997), causality (Silverstein et al., 1998), sequential patterns (Agrawal and Srikant, 1995), episodes (Mannila et al., 1997), multi-dimensional patterns (Lent et al., 1997), max-patterns (Bayardo, 1998), partial periodicity (Han et al., 1999), emerging patterns (Dong and Li,

*The work was supported in part by grants from the Natural Sciences and Engineering Research Council of Canada, and the Networks of Centres of Excellence of Canada (NCE/IRIS-3). A preliminary version of the paper has been published by the authors as Pei et al., "Mining Frequent Itemsets with Convertible Constraints," in the Proceedings of 2001 IEEE International Conference on Data Engineering (ICDE'01), Heidelberg, Germany, April 2001, pp. 433–332.

1999), and many other important data mining tasks. However, frequent pattern mining often generates a very large number of frequent itemsets and rules, which reduces not only the efficiency but also the effectiveness of mining since users have to sift through a large number of mined rules to find useful ones.

Recent work has highlighted the importance of the paradigm of constraint-based mining: the user is allowed to express his focus in mining, by means of a rich class of constraints that capture application semantics. Besides allowing user exploration and control, the paradigm allows many of these constraints to be pushed deep inside mining, confining the search for patterns only to those of interest to the user, and therefore, improving performance. *Metarules* or various kinds of *templates* have been proposed as filters to define the forms of rules to be mined (Klemettinen et al., 1994; Srikant et al., 1997). Itemset constraints have been incorporated into association mining (Srikant et al., 1997). A systematic method for the incorporation of two large classes of constraints, *anti-monotone* and *succinct*, in frequent itemset mining is presented in Ng et al. (1998) and Lakshmanan et al. (1999). A method for mining association rules in large, dense databases by incorporation of user-specified constraints that ensure every mined rule offers a predictive advantage over any of its simplifications, is developed in Bayardo et al. (1999). Constraints specified using regular expressions are investigated for sequential pattern mining in Garofalakis et al. (1999). A systematic study on constraint-based sequential pattern mining is presented in Pei et al. (2002). Constraint-based mining of correlations, by exploration of *anti-monotonicity* and *succinctness*, as well as *monotonicity*, is studied in Grahne et al. (2000).

While previous studies cover a large class of useful constraints, many other useful and natural constraints remain. For example, consider the constraints $avg(S)\theta v$, $median(S)\theta v$, and $sum(S)\theta v$ ($\theta \in \{<, >, \leq, \geq\}$). The first two are neither anti-monotone, nor monotone, nor succinct. The last one is anti-monotone when θ is $\leq$ and *all items have non-negative values*. If S can contain items of arbitrary values, $sum(S) \leq v$ is rather like the first two constraints. Intuitively, this implies that such constraints are hard to optimize. In this paper, we investigate a whole class of constraints that subsumes these examples. The main idea is that certain constraints that exhibit no nice properties in general cases may do so in the presence of certain item ordering. We make the following contributions.

- We introduce the concept of *convertible constraints* and classify them into three classes: *convertible anti-monotone*, *convertible monotone*, and *strongly convertible*. This covers a good number of useful constraints which were previously regarded tough, including all the examples above.
- We characterize the class of convertible constraints using the notion of *prefix monotone* functions, and study the arithmetical closure properties of such functions. As a byproduct, we can show a good class of constraints involving arithmetic are convertible. E.g., we show that $max(S)/avg(S) \leq v$ is convertible anti-monotone and $median(S) - min(S) \geq v$ is convertible monotone.
- We show that convertible constraints cannot be pushed deep into the basic Apriori framework. However, they can be pushed deep into the frequent pattern growth mining. We thus develop algorithms for fast mining of frequent itemsets satisfying the various constraints. We also discuss how multiple convertible constraints can be incorporated in fast frequent pattern mining.

- We report our results from a detailed set of experiments, which show the effectiveness of the algorithms developed.

This study distinguishes itself from the previous works on constraint-based frequent-pattern mining in the following aspects.

- As argued before, the previous works on constraint-based frequent-pattern mining that relied on properties like anti-monotonicity, succinctness, or monotonicity (e.g., Ng et al., 1998; Lakshmanan et al., 1999; Grahne et al., 2000) cannot handle the constraints studied in this paper.
- There have been many studies on constraint-based search algorithms in artificial intelligence, such as Webb (1995) and Rymon (1992). Our study is distinguished from theirs in two aspects: (i) we find the *complete set of frequent itemsets satisfying the constraints*, while their algorithms find *some* feasible solutions satisfying the constraints; and (ii) our goal is to find methods scalable in large databases, while their algorithms are mostly main memory-based.

The remainder of the paper is arranged as follows. Section 2 motivates the problem of frequent itemset mining with constraints. Convertible constraints are proposed and studied in Section 3. The algorithms of mining frequent patterns with convertible constraints are developed in Section 4. The experimental results are reported in Section 5. In Section 6, we discuss how to mine frequent patterns with multiple convertible constraints. Section 7 concludes the paper.

2. Problem definition: Frequent itemset mining with constraints

Let $I = \{i_1, i_2, \ldots, i_m\}$ be a set of all *items*, where an item is an object with some predefined attributes (e.g., price, weight, etc.). A *transaction* $T = \langle tid, I_t \rangle$ is a tuple, where *tid* is the *identifier* of the transaction and $I_t \subseteq I$. A transaction database $\mathcal{T}$ consists of a set of transactions. An itemset $S \subseteq I$ is a subset of the set of items. A k-itemset S is an itemset of size k, i.e., $|S| = k$. As a notational abbreviation, we write itemset as $S = i_{j_1} i_{j_2} \ldots i_{j_k}$, omitting set brackets.

An itemset S is contained in a transaction $T = \langle tid, I_t \rangle$ if and only if $S \subseteq I_t$. The *support* $sup(S)$ of an itemset S in a transaction database $\mathcal{T}$ is the number of transactions in $\mathcal{T}$ containing S. Given a *support threshold* ξ ($1 \leq \xi \leq |\mathcal{T}|$), an itemset S is *frequent* provided $sup(S) \geq \xi$.

A *constraint* C is a predicate on the powerset of the set of items I, i.e., $C : 2^I \rightarrow \{true, false\}$. An itemset S satisfies a constraint C if and only if $C(S)$ is true. The set of itemsets satisfying a constraint C is $\text{SAT}_C(I) = \{S \mid S \subseteq I \wedge C(S) = \text{ true}\}$. We call an itemset in $\text{SAT}_C(I)$ *valid*.

Problem definition. Given a transaction database $\mathcal{T}$, a support threshold ξ, and a set of constraints $\mathcal{C}$, the problem of *mining frequent itemsets with constraints* is to find the

complete set of frequent itemsets satisfying $\mathcal{C}$, i.e., find $\mathcal{F}_{\mathcal{C}} = \{S \mid S \in \text{SAT}_C(I) \wedge sup(S) \geq \xi\}$.

Many kinds of constraints can be associated with frequent itemset mining. Two categories of constraints, *succinctness* and *anti-monotonicity*, were proposed in Ng et al. (1998) and Lakshmanan et al. (1999); whereas the third category, *monotonicity*, was studied in Brin et al. (1997), Grahne et al. (2000) and Pei and Han (2000) in the contexts of mining correlated sets and frequent itemsets. We briefly recall these notions below.

Definition 2.1 (Anti-monotone, Monotone, and Succinct Constraints). A constraint C_a is *anti-monotone* if and only if whenever an itemset S violates C_a, so does any superset of S. A constraint C_m is *monotone* if and only if whenever an itemset S satisfies C_m, so does any superset of S. Succinctness is defined in steps, as follows.

- An itemset $I_s \subseteq I$ is a succinct set, if it can be expressed as $\sigma_p(I)$ for some selection predicate p, where σ is the selection operator.
- $SP \subseteq 2^I$ is a succinct powerset, if there is a fixed number of succinct sets $I_1, I_2, \ldots, I_k \subseteq I$, such that SP can be expressed in terms of the strict powersets of $I_1, \ldots, I_k$ using union and minus.
- Finally, a constraint C_s is *succinct* provided $\text{SAT}_{C_s}(I)$ is a succinct powerset.

We can show the following result.

Theorem 2.1. *Every succinct constraint involving only aggregate functions can be expressed using conjunction and/or disjunction of monotone and anti-monotone constraints.*

Proof Sketch: The proof of the theorem is constructed by induction on the structure of $SAT_C(I)$ of the succinct constraint, according to the definition of succinctness. Here, we consider four essential cases as examples.

- If $SAT_C(I) = 2^{I_c}$, where I_c is a set, then C is an anti-monotone constraint since any pattern satisfying the constraint must be a subset of I_c.
- If $SAT_C(I) = 2^{I_{c_1}} \cup 2^{I_{c_2}}$, then C can be expressed in terms of $C = C_1 \vee C_2$, where C_1 and C_2 are corresponding anti-monotone constraints.
- If $SAT_C(I) = 2^{I_1} - 2^{I_2}$, then the constraint can be expressed as the conjunction of two constraints, $C = C_a \wedge C_m$, where C_a is the anti-monotone constraint corresponding to 2^{I_1}, and C_m is a monotone constraint $S \cap (I_1 - I_2) \neq \emptyset$.
- Especially, if $SAT_C(I) = 2^I - 2^{I_1} - \cdots - 2^{I_m}$, then C is a monotone constraint $S \cap (I - I_1 - \cdots - I_m) \neq \emptyset$. □

These three categories of constraints cover a large class of popularly encountered constraints. A representative subset of commonly used, SQL-based constraints is listed in Table 1.[1] However, there are still many useful constraints, such as $avg(S)\theta v$ and $sum(S)\theta v$ where $\theta \in \{\leq, \geq\}$ (shown in the table) that belong to *none* of the three classes.

Table 1. Characterization of commonly used, SQL-based constraints.

Constraint	Anti-monotone	Monotone	Succinct
$min(S) \leq v$	No	Yes	Yes
$min(S) \geq v$	Yes	No	Yes
$max(S) \leq v$	Yes	No	Yes
$max(S) \geq v$	No	Yes	Yes
$count(S) \leq v$	Yes	No	Weakly
$count(S) \geq v$	No	Yes	Weakly
$sum(S) \leq v\ (\forall a \in S, a \geq 0)$	Yes	No	No
$sum(S) \geq v\ (\forall a \in S, a \geq 0)$	No	Yes	No
$sum(S)\,\theta\, v, \theta \in \{\leq, \geq\}\ (\forall a \in S, a\,\theta\, 0)$	No	No	No
$range(S) \leq v$	Yes	No	No
$range(S) \geq v$	No	Yes	No
$avg(S)\,\theta\, v, \theta \in \{\leq, \geq\}$	No	No	No
$sup(S) \geq \xi$	Yes	No	No
$sup(S) \leq \xi$	No	Yes	No

Example 1. Let Table 2 be our running transaction database $\mathcal{T}$, with a set of items $I = \{a, b, c, d, e, f, g, h\}$.

Let the support threshold be $\xi = 2$. Itemset $S = acd$ is frequent since it is in transactions 10 and 30, respectively. The complete set of frequent itemsets are listed in Table 3.

Let each item have an attribute *value* (such as *profit*), with the concrete value shown in Table 4. In all constraints such as $sum(S)\theta v$, we implicitly refer to this value.

The constraint $range(S) \leq 15$ requires that for an itemset S, the value range of the items in S must be no greater than 15. It is an anti-monotone constraint, in the sense that if an itemset, say ab, violates the constraint, any of its supersets will violate it; and thus ab can be removed safely from the candidate set during an Apriori-like frequent itemset mining process (Ng et al., 1998). However, the constraint $C_{avg} \equiv avg(S) \geq 25$ is not anti-monotone (nor monotone, nor succinct, which can be verified by readers). For example, $avg(df) = (10 + 30)/2 < 25$, violates the constraint. However, upon adding one more item a, $avg(adf) = (40 + 10 + 30)/3 \geq 25$, adf satisfies C_{avg}.

Table 2. The transaction database $\mathcal{T}$ in Example 1.

Transaction ID	Items in transaction
10	a, b, c, d, f
20	b, c, d, f, g, h
30	a, c, d, e, f
40	c, e, f, g

Table 3. Frequent itemsets with support threshold $\xi = 2$ in transaction database $\mathcal{T}$ in Table 2.

Length l	Frequent l-itemsets
1	a, b, c, d, e, f, g
2	$ac, ad, af, bc, bd, bf, cd, ce, cf, cg, df, ef, fg$
3	$acd, acf, adf, bcd, bcf, bdf, cdf, cef, cfg$
4	$acdf, bcdf$

Table 4. The values (such as profit) of items in Example 1.

Item	Value
a	40
b	0
c	−20
d	10
e	−30
f	30
g	20
h	−10

This example scratches the surface of a large class of useful constraints involving *avg*, *median*, etc. as well as arithmetic. Exploiting them in mining calls for new techniques, which is the subject of this paper.

3. Convertible constraints and their classification

Before introducing the concept of convertible constraint, we motivate it with an example.

Example 2. Suppose we wish to mine frequent itemsets over transaction database $\mathcal{T}$ in Table 2, with the support threshold $\xi = 2$ and with constraint $C \equiv avg(S) \geq 25$,

The complete set of frequent itemsets satisfying C can be obtained by first mining the frequent itemsets without using the constraint (i.e., Table 3) and then filtering out those not satisfying the constraint. Since the constraint is neither anti-monotone, nor monotone, nor succinct, it cannot be directly incorporated into an Apriori-style algorithm. E.g., itemset fg satisfies the constraint, while its subset g and its superset dfg do not.

If we arrange the items in *value-descending* order, $\langle a, f, g, d, b, h, c, e\rangle$, we can observe an interesting property, as follows. Writing itemsets w.r.t. this order leads to a notion of a prefix. E.g., afd has af and a as its prefixes. Interestingly, the average of an itemset is no more than that of its prefix, according to this order.

3.1. Convertible constraints

The observation made in Example 2 motivates the following definition. We will frequently make use of an order[2] over the set of all the items and assume itemsets are written according to this order.

Definition 3.1 (Prefix itemset). Given an order $\mathcal{R}$ over the set of items I, an itemset $S' = i_1 i_2 \dots i_l$ is called a *prefix* of itemset $S = i_1 i_2 \dots i_m$ w.r.t. $\mathcal{R}$, where items in both itemsets are listed according to order $\mathcal{R}$ and $(l \leq m)$. S' is called a *proper prefix* of S if $(l < m)$.

We next formalize convertible constraints as follows.

Definition 3.2 (Convertible constraints). A constraint C is *convertible anti-monotone* provided there is an order $\mathcal{R}$ on items such that whenever an itemset S satisfies C, so does any prefix of S. It is *convertible monotone* provided there is an order $\mathcal{R}$ on items such that whenever an itemset S violates C, so does any prefix of S. A constraint is *convertible* whenever it is convertible anti-monotone or monotone.

Note that any anti-monotone (resp., monotone) constraint is trivially convertible anti-monotone (resp., convertible monotone): just pick any order on items.

Example 3. We show $avg(S)\theta v$ where $\theta \in \{\leq, \geq\}$ is a convertible constraint.

Let $\mathcal{R}$ be the value-descending order. Given an itemset $S = a_1 a_2 \dots a_l$ satisfying the constraint $avg(S) \geq v$, where items in S are listed in the order $\mathcal{R}$. For each prefix $S' = a_1 \dots a_k$ of S $(1 \leq k \leq l)$, since $a_k \geq a_{k+1} \geq \cdots \geq a_{l-1} \geq a_l$, we have $avg(S') \geq avg(S' \cup \{a_{k+1}\}) \geq \cdots \geq avg(S) \geq v$. This implies S' also satisfies the constraint. So, constraint $avg(S) \geq v$ is convertible anti-monotone. Similarly, it can be shown that constraint $avg(S) \leq v$ is convertible monotone.

Interestingly, if the order $\mathcal{R}^{-1}$ (i.e., the reversed order of $\mathcal{R}$) is used, the constraint $avg(S) \geq v$ can be shown convertible monotone. We leave this as an exercise to the reader.

In summary, constraint $avg(S)\theta v$ is convertible constraint. Furthermore, there exists an order $\mathcal{R}$ such that the constraint is convertible anti-monotone w.r.t. $\mathcal{R}$ and convertible monotone w.r.t. $\mathcal{R}^{-1}$.

As another example, let us examine the constraints with function $sum(S)$.

Example 4. As shown in Table 1, constraint $sum(S) \leq v$ is anti-monotone if items are all with non-negative values. However, if items are with negative, zero or positive values, the constraint becomes neither anti-monotone, nor monotone, nor succinct.

Interestingly, this constraint exhibits a "piecewise" convertible monotone or anti-monotone behavior. If $v \geq 0$ in the constraint, the constraint is convertible anti-monotone w.r.t. item value *ascending* order. Given an itemset $S = a_1 a_2 \dots a_l$ such that $sum(S) \leq v$, where items are listed in value ascending order. For a prefix $S' = a_1 a_2 \dots a_j$ $(1 \leq j \leq l)$, if $a_j \leq 0$, that means $a_1 \leq a_2 \leq \cdots \leq a_{j-1} \leq a_j \leq 0$. So, $sum(S') \leq 0 \leq v$. On

the other hand, if $a_j > 0$, we have $0 < a_j \leq a_{j+1} \leq \cdots \leq a_l$. Thus, $sum(S') = sum(S) - sum(a_{j+1} \ldots a_l) < v$. Therefore, $sum(S') \leq v$ in both cases, which means S' satisfies the constraint.

If $v \leq 0$ in the constraint, it becomes convertible monotone w.r.t. item value descending order. We leave it to the reader to verify this.

Similarly, we can also show that, if items are with negative, zero or positive values, constraint $sum(S) \geq v$ is convertible monotone w.r.t. value ascending order when $v \geq 0$, and convertible anti-monotone w.r.t. value descending order when $v \leq 0$. □

The following lemma can be proved with a straightforward induction.

Lemma 3.1. *Let C be a constraint over a set of items I.*
1. *C is convertible anti-monotone if and only if there exists an order $\mathcal{R}$ over I such that for every itemset S and item $a \in I$ such that $\forall x \in S, x \mathcal{R} a$, $C(S \cup \{a\})$ implies $C(S)$.*
2. *C is convertible monotone if and only if there exists an order $\mathcal{R}$ over I such that for every itemset S and item $a \in I$ such that $\forall x \in S, x \mathcal{R} a$, $C(S)$ implies $C(S \cup \{a\})$.*

Proof: We show the first part of the lemma. The second part can be shown similarly.

$\Rightarrow$ (if part) Suppose constraint C has the property that for every itemset S and item $a \in I$ such that item $\forall x \in S, x \mathcal{R} a$, $C(S \cup \{a\})$ implies $C(S)$. For an itemset $S = a_1 a_2 \ldots a_m$ and its prefix $S' = a_1 a_2 \ldots a_l$ $(l \leq m)$, let S_k be itemset $a_1 a_2 \ldots a_k$. $C(S) = C(S_{m-1} \cup \{a_m\}) = true$ implies $C(S_{m-1}) = true$. By induction, we can show that $C(S') = C(S_l) = true$. Thus, C is convertible anti-monotone.

$\Leftarrow$ (only-if part) Given a convertible anti-monotone constraint C, following the definition of convertible anti-monotonicity, the property holds that for every itemset S and item $a \in I$ such that item $\forall x \in S, x \mathcal{R} a$, $C(S \cup \{a\})$ implies $C(S)$. □

The notion of prefix monotone functions, introduced below, is helpful in determining the class of a constraint. We denote the set of real numbers as **R**.

Definition 3.3 (Prefix monotone functions). Given an order $\mathcal{R}$ over a set of items I, a function $f : 2^I \rightarrow \mathbf{R}$ is a *prefix* (*monotonically*) *increasing function* w.r.t. $\mathcal{R}$ if and only if for every itemset S and its prefix S' w.r.t. $\mathcal{R}$, $f(S') \leq f(S)$. A function $g : 2^I \rightarrow \mathbf{R}$ is called a *prefix* (*monotonically*) *decreasing function* w.r.t. $\mathcal{R}$ if and only if for every itemset S and its prefix S' w.r.t. $\mathcal{R}$, $g(S') \geq g(S)$.

We have the following lemma on the determination of prefix monotone functions. The proof is similar to that of Lemma 3.1.

Lemma 3.2. *Given an order $\mathcal{R}$ over a set of items I,*
1. *a function $f : 2^I \rightarrow \mathbf{R}$ is a prefix decreasing function w.r.t. $\mathcal{R}$ if and only if for every itemset S and item a such that $\forall x \in S, x \mathcal{R} a$, $f(S) \geq f(S \cup \{a\})$.*
2. *A function $g : 2^I \rightarrow \mathbf{R}$ is a prefix increasing function w.r.t. $\mathcal{R}$ if and only if for every itemset S and item a such that $\forall x \in S, x \mathcal{R} a$, $g(S) \leq g(S \cup \{a\})$.*

Proof: We show the first part of the lemma. The second part can be proved similarly.

$\Leftarrow$ Let $f : 2^I \rightarrow \mathbf{R}$ be a prefix decreasing function. Itemset S is a prefix of $S \cup \{a\}$ if $\forall x \in S, x \mathcal{R} a$. According to the definition of prefix decreasing function, we have $f(S) \geq f(S \cup \{a\})$.

$\Rightarrow$ Suppose function $f : 2^I \rightarrow \mathbf{R}$ has the property that for every itemset S and item a such that $\forall x \in S, x \mathcal{R} a$, $f(S) \geq f(S \cup \{a\})$. For an itemset $S = a_1 a_2 \ldots a_m$ and its prefix $S' = a_1 a_2 \ldots a_l$ $(l \leq m)$, we have $f(S') = f(a_1 a_2 \ldots a_l) \leq f(a_1 a_2 \ldots a_l a_{l+1}) \leq \cdots \leq f(a_1 a_2 \ldots a_m) = f(S)$. So, f is a prefix decreasing function. □

It turns out that prefix monotone functions satisfy interesting closure properties with arithmetic. An understanding of this would shed light on characterizing a whole class of convertible functions involving arithmetic. The following theorem establishes the arithmetical closure properties of prefix monotone functions. We say a function $f : 2^I \rightarrow \mathbf{R}$ is positive, provided $\forall S \subseteq I : f(S) > 0$.

Theorem 3.1. *Let f and f' be prefix decreasing functions, and g and g' be prefix increasing functions w.r.t. an order $\mathcal{R}$, respectively. Let c be a positive real number.*

1. *Functions $-f(S)$, $\frac{1}{f(S)}$, $c \cdot g(S)$ and $g(S) + g'(S)$ are prefix increasing functions. Functions $-g(S)$, $\frac{1}{g(S)}$, $c \cdot f(S)$ and $f(S) + f'(S)$ are prefix decreasing functions.*
2. *If f and g are positive functions, then $f(S) \times f'(S)$ is prefix decreasing, and $g(S) \times g'(S)$ is prefix increasing.*
3. *A constraint $h(S) \geq v$ (resp., $h(S) \leq v$) is convertible anti-monotone (resp., monotone) if and only if h is prefix decreasing. Similarly, $h(S) \geq v$ (resp., $h(S) \leq v$) is convertible monotone (resp., anti-monotone) if and only if h is prefix increasing.*

Proof: The theorem follows related definitions immediately. □

Example 5. As an illustration, notice that $avg(S)$ is a prefix decreasing function w.r.t. value-descending order, and $avg(S) \geq 20$ is convertible anti-monotone w.r.t. the same order. Also, $\max(S)$ is a prefix increasing[3] function w.r.t. this order. From Theorem 3.1, it follows that $1/\text{avg}(S)$ is prefix increasing and hence $max(S)/avg(S)$ is prefix increasing.[4] Consequently, we immediately deduce that $max(S)/avg(S) \leq v$ is convertible anti-monotone w.r.t. this order.

We know from Theorem 2.1 that a succinct constraint can be expressed in terms of conjunction and/or disjunction of anti-monotone and monotone constraints. By definition, every monotone/anti-monotone is convertibly so. A natural question is, what is the relationship between succinct constraints and convertible constraints? The following theorem settles this question.

Theorem 3.2. *Every succinct constraint is either anti-monotone, or monotone, or convertible.*

Proof Sketch: The proof of the theorem is constructed by induction on the structure of $\text{SAT}_C(I)$ of a succinct constraint C, according to the definition of succinctness.

Suppose C is a succinct constraint over I, the set of items.

- $\mathcal{F}_C(I) = 2^{I_C}$, where $I_C \subseteq I$. As shown in Theorem 2.1, constraint C is anti-monotone. Interestingly, C is also convertible monotone w.r.t. order $\mathcal{R}$, where $\forall a \in I_C, b \in I - I_C$, $b\,\mathcal{R}\,c$.
- $\mathcal{F}_C(I) = 2^{I_1} \cup 2^{I_2}$, where $I_1, I_2 \subseteq I$. Constraint C is anti-monotone (Theorem 2.1). However, C is not convertible monotone in this case.
- $\mathcal{F}_C(I) = 2^{I_1} - 2^{I_2}$, where $I_1, I_2 \subseteq I$. Constraint C is convertible anti-monotone w.r.t. order $\mathcal{R}$, where $\forall a \in I_1 - I_2$, $b \in I - (I_1 - I_2)$, $a\mathcal{R}b$. Please note that C is also convertible monotone w.r.t. $\mathcal{R}^{-1}$.
- Especially, $\mathcal{F}_C(I) = 2^I - 2^{I_1} - \cdots - 2^{I_m}$, where $I_1, \ldots, I_m \subseteq I$. Constraint C is monotone (Theorem 2.1). □

As an example, consider a succinct constraint C whose solution space $\text{SAT}_C(I)$ is described as $2^{I_1} - 2^{I_2}$, where $I_1, I_2 \subseteq I$, and $I_i = \sigma_{p_i}(I)$, p_i being a selection predicate, $i = 1, 2$. Consider an order $\mathcal{R}$ such that all the items in $I_1 - I_2$ come before any item in $I - (I_1 - I_2)$, but otherwise the items are ordered arbitrarily. Then, it is easy to see that w.r.t. $\mathcal{R}$, C is convertible anti-monotone and w.r.t. $\mathcal{R}^{-1}$, it is convertible monotone.

3.2. Strongly convertible constraint

Some convertible constraints have the additional desirable property that w.r.t. an order $\mathcal{R}$ they are convertible anti-monotone, while w.r.t. its inverse $\mathcal{R}^{-1}$ they are convertible monotone. E.g., $avg(S) \leq v$ is convertible monotone w.r.t. value ascending order and convertible anti-monotone w.r.t. value descending order (see also Example 3). This property provides great flexibility in data mining query optimization.

Definition 3.4 (Strongly convertible constraint). A constraint C_{sc} is called a *strongly convertible constraint*, provided there exists an order $\mathcal{R}$ over the set of items such that C_{sc} is convertible anti-monotone w.r.t. $\mathcal{R}$ and convertible monotone w.r.t. $\mathcal{R}^{-1}$.

Notice that $median(S)\theta v$ ($\theta \in \{\leq, \geq\}$) is also strongly convertible. Clearly, not every convertible constraint is strongly convertible. E.g., $max(S)/avg(S) \leq v^5$ is convertible anti-monotone w.r.t. value descending order, when all the items have a non-negative value. However, it is not convertible monotone w.r.t. value ascending order.

The following lemma links strongly convertible constraints to prefix monotone functions.

Lemma 3.3. *Constraint $f(S)\theta v$ is strongly convertible, if and only if there exists an order $\mathcal{R}$ over the set of items such that f is a prefix decreasing function w.r.t. $\mathcal{R}$ and a prefix increasing function w.r.t. $\mathcal{R}^{-1}$.*

Proof: The lemma follows Theorem 3.1 immediately. □

For example, $avg(S)$ and $median(S)$ are both prefix decreasing w.r.t. value descending order and prefix increasing w.r.t. value ascending order.

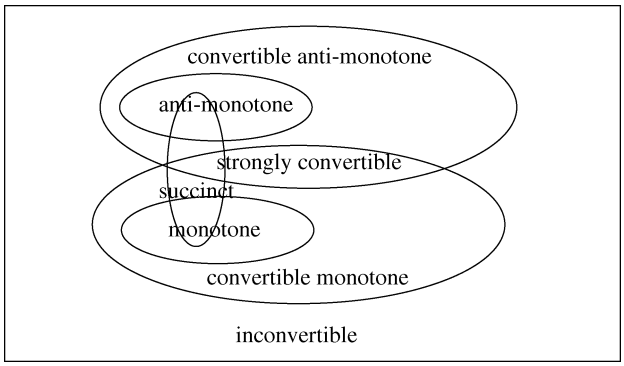

Figure 1. A classification of constraints and their relationships.

There still exist some constraints that cannot be pushed by item ordering. For example, the constraint $avg(S) - median(S) = 0$, which requires that the median item in the itemset is equal to the average value, does not admit any natural ordering on items w.r.t. which it is convertible. We call such constraints *inconvertible*.

3.3. Summary: A classification on constraints

As a general picture, constraints (only involving aggregate functions) can be classified into the following categories according to their interactions with the frequent itemset mining process: *anti-monotone, monotone, succinct* and *convertible*, which in turn can be subdivided into *convertible anti-monotone* and *convertible monotone*. The intersection of the last two categories is precisely the class of *strongly convertible* constraints (which can be treated either as convertible anti-monotone or monotone by ordering the items properly). Figure 1 shows the relationship among the various classes of constraints.

Some commonly used convertible constraints are listed in Table 5.

4. Mining algorithms

In this section, we explore how to mine frequent itemsets with convertible constraints efficiently. The general idea is to push the constraint into the mining process as deep as possible, thereby pruning the search space.

In Section 4.1, we first argue that the Apriori algorithm cannot be extended to mining with convertible constraints efficiently. Then, a new method is proposed by examining an example. Section 4.2 presents the algorithm $\mathcal{FIC}^{\mathcal{A}}$ for mining frequent itemsets with convertible anti-monotone constraints. Algorithm $\mathcal{FIC}^{\mathcal{M}}$, which computes the complete set of frequent itemsets with convertible monotone constraint, is given in Section 4.3. Section 4.4 discusses mining frequent itemsets with strongly convertible constraints.

Table 5. Characterization of some commonly used, SQL-based convertible constraints. ($*$ means it depends on the specific constraint).

Constraint	Convertible anti-monotone	Convertible monotone	Strongly convertible
$avg(S)\theta v$ ($\theta \in \{\leq, \geq\}$)	Yes	Yes	Yes
$median(S)\theta v$ ($\theta \in \{\leq, \geq\}$)	Yes	Yes	Yes
$sum(S) \leq v$ ($v \geq 0, \forall a \in S, a\vartheta 0, \theta, \vartheta \in \{\leq, \geq\}$)	Yes	No	No
$sum(S) \leq v$ ($v \leq 0, \forall a \in S, a\vartheta 0, \theta, \vartheta \in \{\leq, \geq\}$)	No	Yes	No
$sum(S) \geq v$ ($v \geq 0, \forall a \in S, a\vartheta 0, \theta, \vartheta \in \{\leq, \geq\}$)	No	Yes	No
$sum(S) \geq v$ ($v \leq 0, \forall a \in S, a\vartheta 0, \theta, \vartheta \in \{\leq, \geq\}$)	Yes	No	No
$f(S) \geq v$ (f is a prefix decreasing function)	Yes	$*$	$*$
$f(S) \geq v$ (f is a prefix increasing function)	$*$	Yes	$*$
$f(S) \leq v$ (f is a prefix decreasing function)	$*$	Yes	$*$
$f(S) \leq v$ (f is a prefix increasing function)	Yes	$*$	$*$

4.1. Mining frequent itemsets with convertible constraints: An example

We first show that convertible constraints cannot be pushed deep into the Apriori-like mining.

Remark 4.1. A convertible constraint that is neither monotone, nor anti-monotone, nor succinct, cannot be pushed deep into the Apriori mining algorithm.

Rationale. As observed earlier for such a constraint (e.g., $avg(S) \leq v$), subsets (supersets) of a valid itemset could well be invalid and vice versa. Thus, within the levelwise framework, no direct pruning based on such a constraint can be made. In particular, whenever an invalid subset is eliminated without support counting, its supersets that are not suffixes cannot be pruned using frequency.

For example, itemset df in our running example violates the constraint $avg(S) \geq 25$. However, an Apriori-like algorithm cannot prune such itemsets. Otherwise, its superset adf, which satisfies the constraint, cannot be generated.

Before giving our algorithms for mining with convertible constraints, we give an overview in the following example.

Example 6. Let us mine frequent itemsets with constraint $C \equiv avg(S) \geq 25$ over transaction database $\mathcal{T}$ in Table 2, with the support threshold $\xi = 2$. Items in every itemset are listed in value descending order $\mathcal{R}$: $\langle a(40), f(30), g(20), d(10), b(0), h(-10), c(-20), e(-30)\rangle$. It is shown that constraint C is *convertible anti-monotone* w.r.t. $\mathcal{R}$. The mining process is shown in figure 2.

By scanning $\mathcal{T}$ once, we find the support counts for every item. Since h appears in only one transaction, it is an infrequent items and is thus dropped without further consideration. The set of frequent 1-itemsets are a, f, g, d, b, c and e, listed in order $\mathcal{R}$. Among them, only

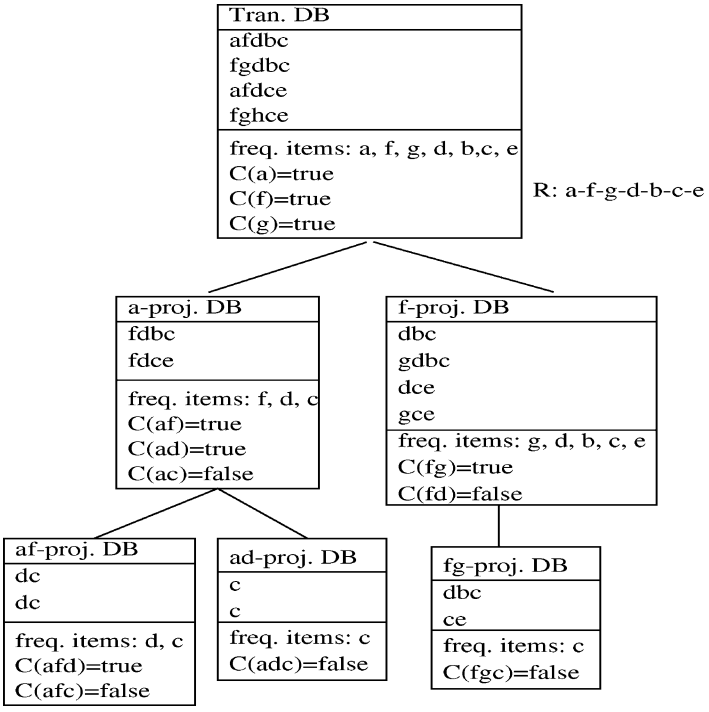

Figure 2. Mining frequent itemsets satisfying constraint $avg(S) \geq 25$.

a and f satisfy the constraint.[6] Since C is a convertible anti-monotone constraint, itemsets having g, d, b, c or e as prefix cannot satisfy the constraint. Therefore, the set of frequent itemsets satisfying the constraint can be partitioned into two subsets:

1. The ones having itemset a as a prefix w.r.t. $\mathcal{R}$, i.e., those containing item a; and
2. The ones having itemset f as a prefix w.r.t. $\mathcal{R}$, i.e., those containing item f but no a.

The two subsets form two projected databases (Han et al., 2000) which are mined respectively.

1. Find frequent itemsets satisfying the constraint and having a as a prefix. First, a is a frequent itemset satisfying the constraint. Then, the frequent itemsets having a as a proper prefix can be found in the subset of transactions containing a, which is called *a-projected database*. Since a appears in every transaction in the a-projected database, it is omitted. The a-projected database contains two transactions: $bcdf$ and $cdef$. Since items b and e are infrequent within this projected database, neither ab nor ae can be frequent. So, they are pruned. The frequent items in the a-projected database is f, d, c, listed in the order $\mathcal{R}$. Since ac does not satisfy the constraint, there is no need to create an ac-projected database.

To check what can be mined in the a-projected database with af and ad, as prefix, respectively, we need to construct the two projected databases and mine them. This process is similar to the mining of a-projected databases. The *af-projected database* contains two frequent items d and c, and only afd satisfies the constraint. Moreover, since $afdc$ does not satisfy the constraint, the process in this branch is complete. Since afc violates the constraint, there is no need to construct afc-projected database. The *ad-projected database* contains one frequent item c, but adc does not satisfy the constraint. Therefore, the set of frequent itemsets satisfying the constraint and having a as prefix contains a, af, afd, and ad.

2. Find frequent itemsets satisfying the constraint and having f as a prefix. Similarly, the *f-projected database* is the subset of transactions containing f, with both a and f removed. It has four transactions: bcd, $bcdg$, cde and ceg. The frequent items in the projected database are g, d, b, c, e, listed in the order of $\mathcal{R}$. Since only itemsets fg and fd satisfy the constraint, we only need to explore if there is any frequent itemset with fg or fd as a proper prefix that satisfies the constraint. The *projected fg-database* contains no frequent itemset with fg as a proper prefix that satisfies the constraint. Since b is the item immediately after d in order $\mathcal{R}$, and fdb violates the constraint, any itemset with fd as a proper prefix cannot satisfy the constraint. Thus, f and fg are the only two frequent itemsets having f as a prefix and satisfying the constraint.

In summary, the complete set of frequent itemsets satisfying the constraint contains 6 itemsets: a, f, af, ad, afd, fg. Our new method generates and tests only a small set of itemsets.

4.2. $\mathcal{FIC}^{\mathcal{A}}$: Mining frequent itemsets with convertible anti-monotone constraint

Now, let us justify the correctness and completeness of the mining process in Example 6.

First, we show that the complete set of frequent itemsets satisfying a given convertible anti-monotone constraint can be partitioned into several non-overlapping subsets. It leads to the soundness of our algorithmic framework.

Lemma 4.1. *Consider a transaction database $\mathcal{T}$, a support threshold ξ and a convertible anti-monotone constraint C w.r.t. an order $\mathcal{R}$ over a set of items I. Let $a_1, a_2, \ldots, a_m$ be the items satisfying C. The complete set of frequent itemsets satisfying C can be partitioned into m disjoint subsets: the jth subset ($1 \leq j \leq m$) contains frequent itemsets satisfying C and having a_j as a prefix.*

Proof Sketch: The lemma follows on showing: (i) every frequent itemset S satisfying C must be in the jth subset, for some j, $1 \leq j \leq m$, and (ii) no two subsets overlap. □

We mine the subsets of frequent itemsets satisfying the constraint by constructing the corresponding *projected database*.

Definition 4.1 (Projected database). Given a transaction database $\mathcal{T}$, an itemset α and an order $\mathcal{R}$.

1. Itemset β is called the *max-prefix projection* of transaction $\langle tid, I_t \rangle \in \mathcal{T}$ w.r.t. $\mathcal{R}$, if and only if (1) $\alpha \subseteq I_t$ and $\beta \subseteq I_t$; (2) α is a prefix of β w.r.t. $\mathcal{R}$; and (3) there exists no proper superset γ of β such that $\gamma \subseteq I_t$ and γ also has α as a prefix w.r.t. $\mathcal{R}$.
2. The *α-projected database* is the collection of max-prefix projections of transactions containing α, w.r.t. $\mathcal{R}$.

Remark 4.2. Given a transaction database $\mathcal{T}$, a support threshold ξ and a convertible anti-monotone constraint C. Let α be a frequent itemset satisfying C. The complete set of frequent itemsets satisfying C and having α as a prefix can be mined from the α-projected database.

Rationale. To mine frequent itemsets having α as a prefix, only the transactions containing α is needed. Furthermore, according to the definition of convertible anti-monotonicity, the information about itemsets having α as a prefix is sufficient to serve the mining with the constraint. That information is completely retained in the max-prefix projections. So we have the lemma.

The mining process can be further improved by the following lemma.

Definition 4.2 (Ascending and descending orders). An order $\mathcal{R}$ over a set of items I is called an *ascending order* for function $h : 2^I \rightarrow \mathbf{R}$ if and only if (1) for items a and b, $h(a) < h(b)$ implies $a\mathcal{R}b$, and (2) for itemsets $\alpha \cup \{a\}$ and $\alpha \cup \{b\}$ such that both of them have α as a prefix and $a\mathcal{R}b$, $f(\alpha \cup \{a\}) \leq f(\alpha \cup \{b\})$. $\mathcal{R}^{-1}$ is called a *descending order* for function h.

For example, it can be verified that the value ascending order is an *ascending* order for function $avg(S)$ and a *descending order* for function $max(S)/avg(S)$.

Lemma 4.2. *Given a convertible anti-monotone constraint $C \equiv f(S)\theta v$ ($\theta \in \{\leq, \geq\}$) w.r.t. ascending/descending order $\mathcal{R}$ over a set of items I, where f is a prefix function. Let α be a frequent itemset satisfying C and $a_1, a_2, \ldots, a_m$ be the set of frequent items in the α-projected database, listed in the order of $\mathcal{R}$.*

1. *If itemset $\alpha \cup \{a_i\}$ ($1 \leq i < m$) violates C, for j such that $i < j \leq m$, itemset $\alpha \cup \{a_j\}$ also violates C.*
2. *If itemset $\alpha \cup \{a_j\}$ ($1 \leq j < m$) satisfies C, but $\alpha \cup \{a_j, a_{j+1}\}$ violates C, no frequent itemset having $\alpha \cup \{a_j\}$ as a proper prefix satisfies C.*

Proof: The constraint C must be in one of the two forms: (1) f is a prefix ascending function w.r.t. descending order $\mathcal{R}$ and $C \equiv f(S) \geq v$ or (2) f is a prefix descending function w.r.t. ascending order $\mathcal{R}$ and $C \equiv f(S) \leq v$. Here, we show the lemma holds for the first case. The second case can be shown similarly.

Suppose $f(\alpha \cup \{a_i\}) < v$S. Since $\mathcal{R}$ is a descending order, $a_i\mathcal{R}a_j$ implies $f(\alpha \cup \{a_j\}) \leq f(\alpha \cup \{a_i\})$. That means itemset $\alpha \cup \{a_j\}$ also violates C.

Alternatively, suppose $f(\alpha \cup \{a_j\}) \geq v$ but $f(\alpha \cup \{a_j, a_{j+1}\}) < v$. For any item a_{j+k} after a_j in the order of $\mathcal{R}$, $f(\alpha \cup \{a_j, a_{j+k}\}) \leq f(\alpha \cup \{a_j, a_{j+1}\})$. So, itemset $\alpha \cup \{a_j, a_{j+k}\}$ must also violate the constraint. □

Based on the above reasoning, we have the algorithm $\mathcal{FIC}^{\mathcal{A}}$ as follows for mining Frequent Itemsets with Convertible Anti-monotone constraints.

Algorithm 1. ($\mathcal{FIC}^{\mathcal{A}}$)

Input: a transaction database $\mathcal{T}$, a support threshold ξ and a convertible anti-monotone constraint C w.r.t. an order $\mathcal{R}$ over a set of items I
Output: the complete set of frequent itemsets satisfying the constraint C
Method: Call $fic_A(\emptyset, \mathcal{T})$;

Function $fic_A(\alpha, \mathcal{T}|_\alpha)$

Parameters: α is the itemset as prefix and $\mathcal{T}|_\alpha$ is the α-projected database.
Method:

1. Scan $\mathcal{T}|_\alpha$ once, find frequent items in $\mathcal{T}|_\alpha$. Let I_α be the set of frequent items within $\mathcal{T}|_\alpha$ such that $\forall a \in I_\alpha, C(\alpha \cup \{a\}) = true$.
2. If $I_\alpha = \emptyset$ return, else $\forall a \in I_\alpha$, output $\alpha \cup \{a\}$ as a frequent itemset satisfying the constraint.
3. If C is in form of $f(S)\theta v$ where f is a prefix function and $\theta \in \{\leq, \geq\}$, using Lemma 4.2 to optimize the mining by removing items b from I_α such that there exists no frequent itemset satisfying C and having $\alpha \cup \{b\}$ as a proper prefix.
4. Scan $\mathcal{T}|_\alpha$ once more, $\forall a \in I|_\alpha$, generate $\alpha \cup \{a\}$-projected database $\mathcal{T}|_{\alpha\cup\{a\}}$.
5. For each item a in $I|_\alpha$, call $fic_A(\alpha \cup \{a\}, \mathcal{T}|_{\alpha\cup\{a\}})$.

Rationale. The correctness and completeness of the algorithm has been reasoned step-by-step in this section. The efficiency of the algorithm is at that it pushes the constraint deep into the mining process, so that we do not need to generate the complete set of frequent itemsets in most cases. Only related frequent itemsets are identified and tested. As shown in Example 6 and in the experimental results, the search space is decreased dramatically when the constraint is sharp.

Based on the above reasoning, we have the following theorem.

Theorem 4.1. *Given a transaction database, a support threshold and a convertible constraint,* $\mathcal{FIC}^{\mathcal{A}}$ (*Algorithm* 1) *computes the complete set of frequent itemsets satisfying the constraint without duplication.*

4.3. $\mathcal{FIC}^{\mathcal{M}}$: *Mining frequent itemsets with monotone constraints*

In the last two subsections, an efficient algorithm for mining frequent itemsets with convertible anti-monotone constraints is developed. Under similar spirit, an algorithm for mining frequent itemsets with convertible monotone constraints can also be developed. Instead of giving details of formal reasoning, we illustrate the ideas using an example and then present the algorithm.

Example 7. Let us mine frequent itemsets in transaction database $\mathcal{T}$ in Table 2 with constraint $C \equiv avg(S) \leq 20$. Suppose the support threshold $\xi = 2$. In this example, we use the value descending order $\mathcal{R}$ exactly as is used in Example 6. Constraint C is convertible monotone w.r.t. order $\mathcal{R}$.

After one scan of transaction database $\mathcal{T}$, the set of frequent 1-itemsets is found. Among the 7 frequent 1-itemsets, g, d, b, c and e satisfy the constraint C. According to the definition of convertible monotone constraints, frequent itemset having one of these 5 itemsets as a prefix must also satisfy the constraint. That is, the g-, d-, b-, c- and e-projected databases can be mined without testing constraint C, because adding smaller items will only decrease the value of avg. But a- and f-projected databases should be mined with constraint C testing. However, as soon as its frequent k-itemsets for any k satisfy the constraint, constraint checking will not be needed for further mining of their projected databases.

We present the algorithm $\mathcal{FIC}^{\mathcal{M}}$ for mining frequent itemsets with convertible monotone constraint as follows.

Algorithm 2. ($\mathcal{FIC}^{\mathcal{M}}$)

Input: A transaction database $\mathcal{T}$, a support threshold ξ and a convertible monotone constraint C w.r.t. an order $\mathcal{R}$ over a set of items I.
Output: The complete set of frequent itemsets satisfying the constraint C.
Method: Call $fic_M(\emptyset, \mathcal{T}, 1)$;
Function $fic_M(\alpha, \mathcal{T}|_\alpha, check\_flag)$
Parameters: α is the itemset as prefix, $\mathcal{T}|_\alpha$ is the α-projected database, and *check_flag* is the flag for constraint checking.
Method:

1. Scan $\mathcal{T}|_\alpha$ once, find frequent items in $\mathcal{T}|_\alpha$. If *check_flag* is 1, let I_α^+ be the set of frequent items within $\mathcal{T}|_\alpha$ such that $\forall a \in I_\alpha^+, C(\alpha \cup \{a\}) = true$, and I_α^- be the set of frequent items within $\mathcal{T}|_\alpha$ such that $\forall b \in I_\alpha^-, C(\alpha \cup \{b\}) = false$. If *check_flag* is 0, let I_α^+ be the set of frequent items within $\mathcal{T}|_\alpha$ and I_α^- be $\emptyset$.
2. $\forall a \in I_\alpha^+$, output $\alpha \cup \{a\}$ as a frequent itemset satisfying the constraint.
3. Scan $\mathcal{T}|_\alpha$ once more, $\forall a \in I|_\alpha^+ \cup I|_\alpha^-$, generate $\alpha \cup \{a\}$-projected database $\mathcal{T}|_{\alpha\cup\{a\}}$.
4. For each item a in $I|_\alpha^+$, call $fic_M(\alpha \cup \{a\}, \mathcal{T}|_{\alpha\cup\{a\}}, 0)$; For each item a in $I|_\alpha^-$, call $fic_M(\alpha \cup \{a\}, \mathcal{T}|_{\alpha\cup\{a\}}, 1)$;

Rationale. The correctness and completeness of the algorithm can be shown based on the similar reasoning in Section 4.2. Here, we analyze the difference between $\mathcal{FIC}^{\mathcal{M}}$ with an Apriori-like algorithm using constraint-checking as post-processing.

Both $\mathcal{FIC}^{\mathcal{M}}$ and Apriori-like algorithms have to generate the complete set of frequent itemsets, no matter whether the frequent itemsets satisfy the convertible monotone constraint. The frequent itemsets not satisfying the constraint cannot be pruned. This is the inherent difficulty of convertible monotone constraint.

The advantage of $\mathcal{FIC}^{\mathcal{M}}$ against Apriorix-like algorithms lies in the fact that $\mathcal{FIC}^{\mathcal{M}}$ only tests some of frequent itemsets against the constraint. Once a frequent itemset satisfies

the constraint, it guarantees all of frequent itemsets having it as a prefix also satisfy the constraint. Therefore, all that testing can be saved. An Apriori-like algorithm has to check every frequent itemset against the constraint. In the situation such that constraint testing is costly, such as spatial constraints, the saving over constraint testing could be non-trivial. Exploration of spatial constraints is beyond the scope of this paper.

4.4. Mining frequent itemsets with strongly convertible constraints

The main value of strong convertibility is that the constraint can be treated either as convertible anti-monotone or monotone by choosing an appropriate order. The main point to note in practice is when the constraint has a high selectivity (fewer itemsets satisfy it), converting it into an anti-monotone constraint will yield maximum benefits by search space pruning. When the constraint selectivity is low (and checking it is reasonably expensive), then converting it into a monotone constraint will save considerable effort in constraint checking. The constraint $avg(S) \leq v$ is a classic example.

5. Experimental results

To evaluate the effectiveness and efficiency of the algorithms, we performed an extensive experimental evaluation.

In this section, we report the results on a synthetic transaction database with 100 K transactions and 10 K items. The dataset is generated by the standard procedure described in Agrawal and Srikant (1994). In this dataset, the average transaction size and average maximal potentially frequent itemset size are set to 25 and 20, respectively. The dataset contains a lot of frequent itemsets with various lengths. This dataset is chosen since it is typical in data mining performance study.

The algorithms are implemented in C. All the experiments are performed on a 233 MHz Pentium PC with 128 MB main memory, running Microsoft Windows/NT.

To evaluate the effect of a constraint on mining frequent itemsets, we make use of constraint selectivity, where the *selectivity* δ of a constraint C on mining frequent itemsets over transaction database $\mathcal{T}$ with support threshold ξ is defined as

$$\delta = \frac{\text{\# of frequent itemsets NOT satisfying } C}{\text{\# of frequent itemsets}} \tag{1}$$

Therefore, a constraint with 0% selectivity means every frequent itemset satisfies the constraint, while a constraint with 100% selectivity means that the constraint cannot be satisfied by any frequent itemset. The selectivity measure defined here is consistent with those used in Ng et al. (1998) and Lakshmanan et al. (1999).

To facilitate the mining using projected databases, we employ a data structure called FP-tree in the implementations of $\mathcal{FIC}^{\mathcal{A}}$ and $\mathcal{FIC}^{\mathcal{M}}$. FP-tree is first proposed in Han et al. (2000), and also be adopted by Pei and Han (2000) and Pei et al. (2000). It is a prefix tree structure to record complete and compact information for frequent itemset mining. A transaction database/projected database can be compressed into an FP-tree, while all the

consequent projected databases can be derived from it efficiently. We refer readers to Han et al. (2000) for details about FP-tree and methods for FP-tree-based frequent itemset mining.

Since FP-growth (Han et al., 2000) is the FP-tree-based algorithm mining frequent itemsets and much faster than Apriori, we include it in our experiment. It is thus more interesting to compare the performance among $\mathcal{FIC}^{\mathcal{A}}$, $\mathcal{FIC}^{\mathcal{M}}$, and FP-growth than taking Apriori as the only reference method.

5.1. *Evaluation of* $\mathcal{FIC}^{\mathcal{A}}$

To test the efficiency of $\mathcal{FIC}^{\mathcal{A}}$ w.r.t. constraint selectivity in mining frequent itemsets with convertible anti-monotone constraints, a test is performed over the dataset with support threshold $\xi = 0.1\%$. The result is shown in figure 3. Various settings are used in the constraint for various selectivities.

As shown in figure 3, $\mathcal{FIC}^{\mathcal{A}}$ achieves an almost linear scalability with the constraint selectivity. As the selectivity goes up, i.e., when fewer itemsets satisfying the constraint, $\mathcal{FIC}^{\mathcal{A}}$ cuts more search space, since if there is a frequent itemset s which does not satisfy the constraint, it means all the frequent itemsets with s as a prefix can be pruned.

We also compare the runtime of Apriori and FP-growth in the same figure. Both methods first compute the complete set of frequent itemsets and then use the constraint as a filter. So, their runtime is constant w.r.t. constraint selectivity. However, only when the constraint selectivity is 0%, i.e., when every frequent itemset satisfies the constraint, does $\mathcal{FIC}^{\mathcal{A}}$ need as same runtime as FP-growth. In all other situations, $\mathcal{FIC}^{\mathcal{A}}$ always requires less time.

We also tested the scalability of $\mathcal{FIC}^{\mathcal{A}}$ with support threshold and number of transactions, respectively. The corresponding results are shown in figures 4 and 5. From the figures, we can

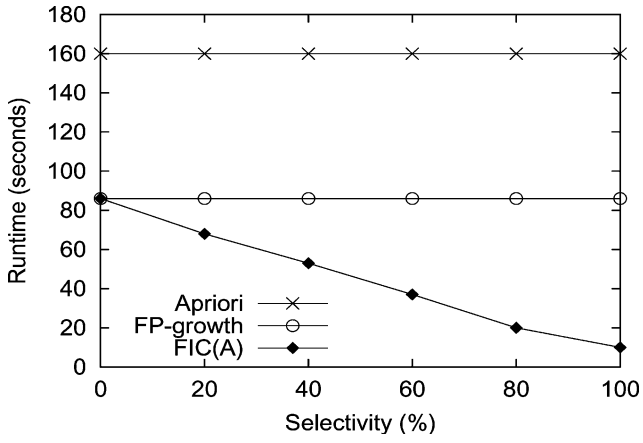

Figure 3. Scalability with constraint selectivity.

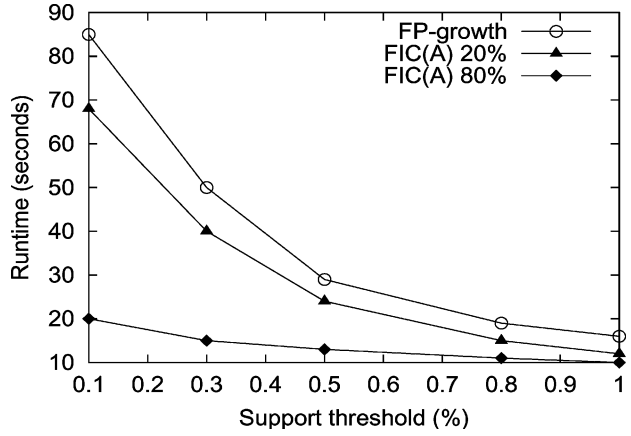

Figure 4. Scalability with support threshold.

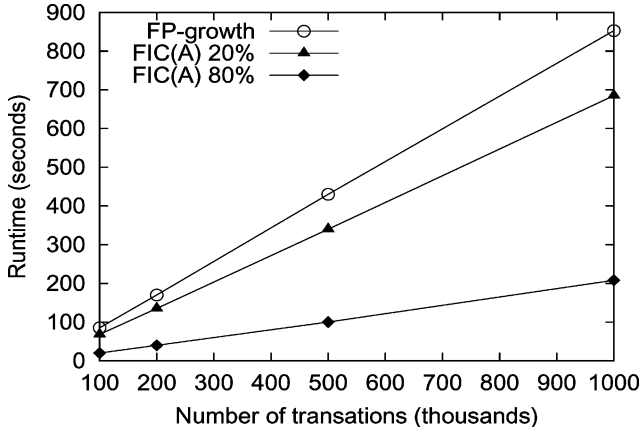

Figure 5. Scalability with number of transactions.

see that $\mathcal{FIC}^{\mathcal{A}}$ is scalable in both cases. Furthermore, the higher the constraint selectivity, the more scalable $\mathcal{FIC}^{\mathcal{A}}$ is. This can be explained by the fact that $\mathcal{FIC}^{\mathcal{A}}$ always cuts more search spaces using constraints with higher selectivity.

5.2. *Evaluation of* $\mathcal{FIC}^{\mathcal{M}}$

As analyzed before, convertible monotone constraint can be used to save the cost of constraint checking, but it cannot cut the search space of frequent itemsets. In our experiments, since we use relatively simple constraints, such as those involving *avg* and *sum*, the cost of constraint checking is CPU-bounded. However, the cost of the whole frequent itemset

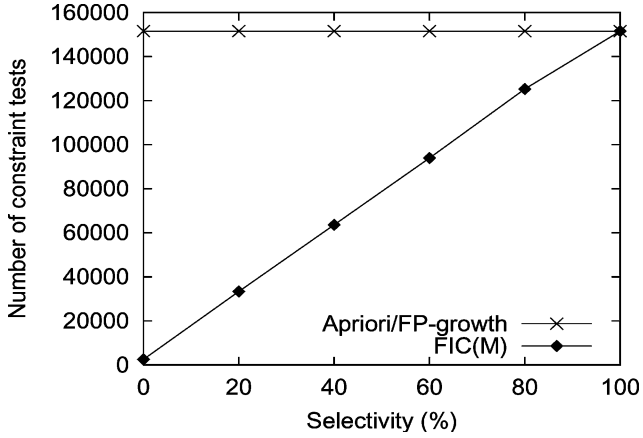

Figure 6. Scalability with constraint selectivity.

mining process is I/O-bounded. This makes the effect of pushing convertible monotone constraint into the mining process hard to be observed from runtime reduction. In our experiments, $\mathcal{FIC}^{\mathcal{M}}$ achieves less than 3% runtime benefit in most cases.

However, if we look at the number of constraint tests performed, the advantage of $\mathcal{FIC}^{\mathcal{M}}$ can be evaluated objectively. $\mathcal{FIC}^{\mathcal{M}}$ can save a lot of effort on constraint testing. Therefore, in the experiments about $\mathcal{FIC}^{\mathcal{M}}$, the number of constraint tests is used as the performance measure.

We test the scalability of $\mathcal{FIC}^{\mathcal{M}}$ with constraint selectivity in mining frequent itemsets with convertible monotone constraint. The result is shown in figure 6, which indicates that $\mathcal{FIC}^{\mathcal{M}}$ has a linear scalability. When the constraint selectivity is low, i.e., most frequent itemsets can pass the constraint checking, most of constraint tests can be saved. This is because once a frequent itemset satisfies a convertible monotone constraint, every subsequent frequent itemset derived from corresponding projected database has that frequent itemset as a prefix and thus satisfies the constraint, too.

We also tested the scalability of $\mathcal{FIC}^{\mathcal{M}}$ with support threshold. The result is shown in figure 7. The figure shows that $\mathcal{FIC}^{\mathcal{M}}$ is scalable. Furthermore, the lower the constraint selectivity, the better the scalability $\mathcal{FIC}^{\mathcal{M}}$ is.

In summary, our experimental results show that the method proposed in this paper is scalable for mining frequent itemsets with convertible constraints in large transaction databases. The experimental results strongly support our theoretical analysis.

6. Discussions: Mining frequent itemsets with multiple convertible constraints

We have studied the push of *single* convertible constraints into frequent itemset mining. "*Can we push multiple constraints deep into the frequent pattern mining process*?"

Multiple constraints in a mining query may belong to the same category (e.g. all are anti-monotone) or to different categories. Moreover, different constraints may be on different

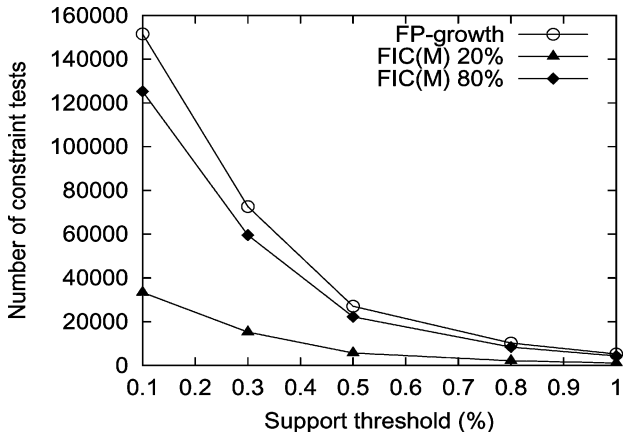

Figure 7. Scalability with support threshold.

properties of items (e.g. some could be on item price, others on sales profits, the number of items, etc.).

As shown in our previous analysis, unlike anti-monotone, monotone and succinct constraints, *convertible* constraints can be mined only by ordering items properly. However, different constraints may require different or even conflicting item ordering. The question is how to deal with this nicely. In the following, we refer to a constraint with a high (low) selectivity as a sharp (blunt) constraint.

In the sequel, we consider mining frequent itemsets with a constraint $C_1 \circ C_2$, where both C_1 and C_2 are convertible constraints, and $\circ \in \{\wedge, \vee\}$.

Case 1. There exists an order $\mathcal{R}$ such that both C_1 and C_2 are convertible w.r.t. $\mathcal{R}$. In such a case, there is no conflict between the two convertible constraints. So, we can push both constraints into the mining process using the order $\mathcal{R}$. We suggest some heuristics as shown in Table 6.

Case 2. There exists a conflict on the order of items. Suppose C_1 requires $\mathcal{R}_1$ and C_2 requires $\mathcal{R}_2$, and $\mathcal{R}_1$ and $\mathcal{R}_2$ is incompatible. In such situations, we should try to satisfy one constraint at first, and then using the order for the other constraint to mine frequent itemsets in the corresponding projected database. The strategies are shown in Table 7.

Interested readers may verify the strategies in Tables 6 and 7 with the similar reasoning as provided in Section 4. We need ways to estimate the selectivities of constraints. In practice, methods such as sampling and business background knowledge often provide useful estimation. Notice that queries may contain an anti-monotone or monotone constraint together with a convertible constraint. Since an anti-monotone or monotone constraint does not impose requirements on item ordering, such a constraint can be treated similarly as Case 1 (Table 6). Also, this discussion can be extended to the cases when there are more than two constraints.

Table 6. Strategies for mining with multiple convertible constraints without conflict on item ordering.

Categories of constraints	$C_1 \vee C_2$	$C_1 \wedge C_2$
Both are convertible anti-monotone	Test the blunt constraint first. Only for itemsets violating both C_1 and C_2, the corresponding projected database can be pruned.	Test the sharp constraint first. For itemsets violating either constraint, their projected database can be pruned.
Both are convertible monotone	Test the blunt constraint first. Once an itemset satisfies either constraint, all the follow-up testing can be waived.	Test the sharp constraint first. Only when an itemset satisfies both constraints, can all the following-up testing be waived.
One is convertible monotone, while the other is convertible anti-monotone	Test the convertible monotone one first. If it is satisfied, the following-up testing can be waived.	Test the convertible anti-monotone constraint first. If it is violated, the corresponding projected database can be pruned. The convertible anti-monotone constraint-checking has to be done all the time, even when the convertible monotone one is satisfied/waived.

Table 7. Strategies for mining with multiple convertible constraints with conflict on item ordering.

Categories of constraints	$C_1 \vee C_2$	$C_1 \wedge C_2$
Both are convertible anti-monotone	Test the blunt constraint, say C_1, first, using order $\mathcal{R}_1$. When a frequent itemset α violates C_1, mine frequent itemsets β in α-projected database, using $\mathcal{R}_2$, such that $\alpha \cup \beta$ satisfies C_2.	Test the sharp constraint, say C_1, using order $\mathcal{R}_1$, all the time. Use C_2 as a post-filter.
Both are convertible monotone	Test the blunt constraint, say C_1, first, using order $\mathcal{R}_1$. When a frequent itemset α violates C_1, mine frequent itemsets β in α-projected database, using $\mathcal{R}_2$, such that $\alpha \cup \beta$ satisfies C_2.	Test the sharp constraint, say C_1, using order $\mathcal{R}_1$ first. When a frequent itemset α satisfies C_1, mine frequent itemsets β in α-projected database, using $\mathcal{R}_2$, such that $\alpha \cup \beta$ satisfies C_2.
One is convertible monotone, while the other is convertible anti-monotone	Test the convertible monotone one, say C_1, first, using $\mathcal{R}_1$. If satisfied, the follow-up testing can be waived. In the α-projected database such that α violates C_1, mine frequent itemsets β using $\mathcal{R}_2$ such that $\alpha \cup \beta$ satisfies C_2.	Test the convertible anti-monotone constraint first. If it is violated, corresponding projected database can be pruned. Use C_2 as a post-filter.

7. Conclusions

Although there have been interesting studies, such as (Ng et al., 1998; Lakshmanan et al., 1999; Grahne et al., 2000), on mining frequent patterns with constraints, constraints involving holistic functions such as *median*, algebraic functions such as *avg*, or even those involving distributive functions like *sum* over sets with positive and negative item values are difficult to incorporate in an optimization process in frequent itemset mining. The reason is such constraints do not exhibit nice properties like monotonicity, etc. A main contribution of this paper is showing that by imposing an appropriate order on items, such tough constraints can be converted into ones that possess monotone behavior. To this end, we made a detailed analysis and classification of the so-called convertible constraints. We characterized them using prefix monotone functions and established their arithmetical closure properties. As a byproduct, we shed light on the overall picture of various classes of constraints that can be optimized in frequent set mining. While convertible constraints cannot be literally incorporated into an Apriori-style algorithm, they can be readily incorporated into the FP-growth algorithm. Our experiments show the effectiveness of the algorithms developed.

We have been working on a systematic implementation of constraint-based frequent pattern mining in a data mining system. More experiments are needed to understand how best to handle multiple constraints. An open issue is given an arbitrary constraint, how can we quickly check if it is (strongly) convertible. We are also exploring the use of constraints in clustering.

Notes

1. For brevity, we show a small subset of representative constraints, involving aggregates. See Ng et al. (1998) and Lakshmanan et al. (1999) for more details.
2. Unless otherwise stated, every order used in this paper is assumed to be total over the set of items.
3. It is also prefix decreasing w.r.t. this order.
4. Assuming all the items have non-negative values.
5. It says the proportion of the max price of any item in the itemset over the average price of the items in the set cannot go over certain limit.
6. The fact that itemset g does not satisfy the constraint implies none of any 1-itemsets after g in order $\mathcal{R}$ can satisfy the constraint *avg*.

References

Agrawal, R. and Srikant, R. 1994. Fast algorithms for mining association rules. In Proc. 1994 Int. Conf. Very Large Data Bases (VLDB'94), Santiago, Chile, pp. 487–499.

Agrawal, R. and Srikant, R. 1995. Mining sequential patterns. In Proc. 1995 Int. Conf. Data Engineering (ICDE'95), Taipei, Taiwan, pp. 3–14.

Bayardo, R.J. 1998. Efficiently mining long patterns from databases. In Proc. 1998 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'98), Seattle, WA, pp. 85–93.

Bayardo, R.J., Agrawal, R., and Gunopulos, D. 1999. Constraint-based rule mining on large, dense data sets. In Proc. 1999 Int. Conf. Data Engineering (ICDE'99), Sydney, Australia, pp. 188–197.

Brin, S., Motwani, R., and Silverstein, C. 1997. Beyond market basket: Generalizing association rules to correlations. In Proc. 1997 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'97), Tucson, Arizona, pp. 265–276.

Dong, G. and Li, J. 1999. Efficient mining of emerging patterns: Discovering trends and differences. In Proc. 1999 Int. Conf. Knowledge Discovery and Data Mining (KDD'99), San Diego, CA, pp. 43–52.

Garofalakis, M., Rastogi, R., and Shim, K. 1999. SPIRIT: Sequential pattern mining with regular expression constraints. In Proc. 1999 Int. Conf. Very Large Data Bases (VLDB'99), Edinburgh, UK, pp. 223–234.

Grahne, G., Lakshmanan, L., and Wang, X. 2000. Efficient mining of constrained correlated sets. In Proc. 2000 Int. Conf. Data Engineering (ICDE'00), San Diego, CA, pp. 512–521.

Han, J., Dong, G., and Yin, Y. 1999. Efficient mining of partial periodic patterns in time series database. In Proc. 1999 Int. Conf. Data Engineering (ICDE'99), Sydney, Australia, pp. 106–115.

Han, J., Pei, J., and Yin, Y. 2000. Mining frequent patterns without candidate generation. In Proc. 2000 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'00), Dallas, TX, pp. 1–12.

Klemettinen, M., Mannila, H., Ronkainen, P., Toivonen, H., and Verkamo, A.I. 1994. Finding interesting rules from large sets of discovered association rules. In Proc. 3rd Int. Conf. Information and Knowledge Management, Gaithersburg, Maryland, pp. 401–408.

Lakshmanan, L.V.S., Ng, R., Han, J., and Pang, A. 1999. Optimization of constrained frequent set queries with 2-variable constraints. In Proc. 1999 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'99), Philadelphia, PA, pp. 157–168.

Lent, B., Swami, A., and Widom, J. 1997. Clustering association rules. In Proc. 1997 Int. Conf. Data Engineering (ICDE'97), Birmingham, England, pp. 220–231.

Mannila, H., Toivonen, H., and Verkamo, A.I. 1994. Efficient algorithms for discovering association rules. In Proc. AAAI'94 Workshop Knowledge Discovery in Databases (KDD'94), Seattle, WA, pp. 181–192.

Mannila, H., Toivonen, H., and Verkamo, A.I. 1997. Discovery of frequent episodes in event sequences. Data Mining and Knowledge Discovery, 1:259–289.

Ng, R., Lakshmanan, L.V.S., Han, J., and Pang, A. 1998. Exploratory mining and pruning optimizations of constrained associations rules. In Proc. 1998 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'98), Seattle, WA, pp. 13–24.

Pei, J. and Han, J. 2000. Can we push more constraints into frequent pattern mining? In Proc. 2000 ACM SIGKDD Int. Conf. Knowledge Discovery in Databases (KDD'00), Boston, MA, pp. 350–354.

Pei, J., Han, J., and Lakshmanan, L.V.S. 2001. Mining frequent itemsets with convertible constraints. In Proc. 2001 Int. Conf. Data Engineering (ICDE'01), Heidelberg, Germany, pp. 433–332.

Pei, J., Han, J., and Mao, R. 2000. CLOSET: An efficient algorithm for mining frequent closed itemsets. In Proc. 2000 ACM-SIGMOD Int. Workshop Data Mining and Knowledge Discovery (DMKD'00), Dallas, TX, pp. 11–20.

Pei, J., Han, J., and Wang, W. 2002. Constraint-based sequential pattern mining in large databases. In Proc. 2002 Int. Conf. on Information and Knowledge Management (CIKM'02), McLean, VA.

Rymon, R. 1992. Search through systematic set enumeration. In Proc. 1992 Int. Conf. Principle of Knowledge Representation and Reasoning (KR'92), Cambridge, MA, pp. 539–550.

Silverstein, C., Brin, S., Motwani, R., and Ullman, J. 1998. Scalable techniques for mining causal structures. In Proc. 1998 Int. Conf. Very Large Data Bases (VLDB'98), New York, NY, pp. 594–605.

Srikant, R., Vu, Q., and Agrawal, R. 1997. Mining association rules with item constraints. In Proc. 1997 Int. Conf. Knowledge Discovery and Data Mining (KDD'97), Newport Beach, CA, pp. 67–73.

Webb, G.I. 1995. Opus: An efficient admissible algorithm for unordered search. Journal of Artificial Intelligence Research, 3:431–465.

Jian Pei received the Bachelor of Engineering and the Master of Engineering degrees, both in Computer Science, from Shanghai Jiao Tong University, China, in 1991 and 1993, respectively, and the Ph.D. degree in Computing Science from Simon Fraser University, Canada, in 2002. He is currently an Assistant Professor of Computer Science and Engineering, and also a participating faculty in the Center of Unified Biometrics and Sensors (CUBS), the State University of New York at Buffalo, USA. His research interests include data mining, data warehousing, online analytical processing, database systems, and bio-informatics. His current research is supported in part by the National Science Foundation (NSF). He has published over 40 research papers in refereed journals, conferences, and workshops. He has served in the program committees and organization committees of over 30 international

conferences and workshops. He is an associate editor of the ACM SIGMOD Digital Symposium Collection, and has been a reviewer for some leading academic journals including ACM and IEEE Transactions. He is a member of the ACM, the ACM SIGMOD, the ACM SIGKDD and the IEEE Computer Society.

Jiawei Han, Professor, Department of Computer Science, University of Illinois at Urbana-Champaign. Previously, he was an Endowed University Professor at Simon Fraser University, Canada. He has been working on research into data mining, data warehousing, database systems, spatial databases, deductive and object-oriented databases, Web databases, bio-medical databases, etc. with over 150 journal and conference publications. He has chaired or served in many program committees of international conferences and workshops, including 2001 and 2002 SIAM-Data Mining Conference (PC co-chair), 2002 International Conference on Data Engineering (PC vice-chair), ACM SIGKDD conferences (2001 best paper award chair, 2002 student award chair), 2002 and 2003 ACM SIGMOD conference (year 2000 demo/exhibit program chair), etc. He also served or is serving on the editorial boards for Data Mining and Knowledge Discovery: An International Journal, IEEE Transactions on Knowledge and Data Engineering, and Journal of Intelligent Information Systems. He has also been serving on the Board of Directors for the Executive Committee of ACM Special Interest Group on Knowledge Discovery and Data Mining (SIGKDD). His textbook "Data Mining: Concepts and Techniques" (Morgan Kaufmann, 2001) has been popularly used for data mining courses in many universities.

Laks V.S. Lakshmanan is a Professor of Computer Science at the University of British Columbia, Vancouver, Canada. He obtained his Bachelor's from the A.C. College of Engineering and Technology, Karaikudi, and his Master's and Ph.D. from the Indian Institute of Science, Bangalore, India. He was a postdoc at the University of Toronto for two years. His research interests span a wide spectrum of topics in Database Systems and related areas, including: relational and object-oriented databases, advanced data models for novel applications, OLAP and data warehousing, database mining, information integration, semi-structured data and XML, directory-enabled networks, and querying the WWW. A common theme underlying his research is to model problems not traditionally viewed as standard database problems and bring database technology to bear on them, thus pushing the frontiers of database technology. He has been a consultant to H.P. Labs, Palo Alto, CA, and AT&T Labs Research, Florham Park, NJ, and a visiting professor at the Limburg Universitair Centrum, Limburg, Belgium, IASI, CNR, Rome, Italy, and the Indian Institute of Science, Bangalore, India. His research is funded by the Natural Sciences and Engineering Research Council of Canada (NSERC), the Network of Centres of Excellence/Institute of Robotics and Intelligent Systems (NCE/IRIS), and Mathematics of Information Technology and Complex Systems (MITACS). Laks has served in the program committees of many top database conferences, including SIGMOD, PODS, SIGKDD, EDBT, and ICDE, and has edited/co-edited several journal special issues. He has published extensively in the standard top database conferences and journals. His honors include a Best Student Paper Award in ICDT, Rome, Italy (1986) and a Best Ph.D. Thesis Gold Medal, IISc, Bangalore (1990). He is currently a Research Fellow of the BC Advanced Systems Institute. Laks serves on the editorial boards of IEEE Transactions on Knowledge and Data Engineering and ACM SIGMOD Digital Symposium Collection (DiSC). Prior to joining UBC, he was a visiting professor at IIT-Bombay and an associate professor at Concordia University, Montreal.