A new concise representation of frequent patterns through disjunctive search space

T. Hamrouni[1,2], I. Denden[1], S. Ben Yahia[1] and E. Mephu Nguifo[2]

[1] Faculty of Sciences of Tunis, Tunis, Tunisia.
{*tarek.hamrouni, sadok.benyahia*}*@fst.rnu.tn*
[2] CRIL-CNRS, IUT de Lens, Lens, France.
{*hamrouni, mephu*}*@cril.univ-artois.fr*

Abstract. The interest in a further pruning of the set of frequent patterns that can be drawn from real-life datasets is growing up. In fact, it is a quite survival reflex towards providing a manageably-sized and reliable knowledge. This fact is witnessed by the proliferation of what is called *concise representation* of frequent patterns. In this paper, we propose an exact concise representation that explores the *disjunctive search space* in addition to the conjunctive one, in contrast with almost all known concise representations which only focussed on the latter space. This representation required the definition of a new disjunctive closure operator. The latter operator partitions the search space into distinct disjunctive equivalence classes and, hence, makes possible to drastically reduce the number of handled patterns. Empirical evidences are presented about the relative size of the new representation *w.r.t.* those based on frequent closed, (closed) non-derivable and essential patterns, respectively.

Keywords: Frequent pattern, Concise representation, Disjunctive search space, Itemset.

1 Introduction and motivations

Within the traditional framework of association rule mining, managing the high number of frequent patterns extracted from real-life datasets becomes an important topic [(1)]. A growing number of works hence explored the conjunctive search space to get out a nucleus of patterns, from which the remaining ones can be derived without information loss. Such an exploration was mainly motivated by the fact that the conjunctive operator – linking items – got the monopoly since the application of association rules in market basket analysis. Such a nucleus is better known as *exact concise representation*. Beyond expected high compactness rates, an exact concise representation should make possible to guess the frequency status of a pattern, and then to exactly retrieve its support when it is frequent enough. The main exact concise representations proposed are those based on frequent closed [1], non-derivable [2], closed non-derivable [3] [(2)] and essential patterns [4]. The first three representations also have the interesting property of being *true* (also called *perfect* in [4]) covers of frequent patterns, since their cardinality is always smaller than that of the frequent pattern set.

[1] Here we are mainly interested in itemsets as a pattern class.

[2] This representation simply gathers the set of closures of frequent non-derivable patterns. It is, hence, smaller in size terms than the previous two ones.

The main originality of the concise representation based on frequent essential patterns stands in the fact that it mainly explores the *disjunctive search space* where elements are characterized by their respective disjunctive supports, instead of conjunctive ones. It hence makes use of the inclusion-exclusion identities [5] to bridge both conjunctive and disjunctive search spaces. Nevertheless, in spite of such originality, this representation suffers from two major disadvantages:
1. It is not self-contained in the sense that the essential pattern set does not make possible by itself to decide whether a pattern is frequent or not. Hence, such a set has to be burdened by additional elements belonging to the positive border of the order ideal induced by the frequency constraint.
2. Several essential patterns can characterize the same set of objects and, therefore, they present a certain form of redundancy.
In this situation, finding a closure operator related to essential patterns would be of paramount importance to get a more reduced concise representation. Indeed, thanks to this operator, many essential patterns will be mapped into the same element within the disjunctive search space. Thus, the obtained representation will be more compact, especially for dense datasets. Furthermore, the simultaneous use of essential patterns and disjunctive closed ones can also ease the detection of their respective disjunctive equivalence classes and, hence, the traversal of the disjunctive search space. This can intensively be explored in many applications as done within the conjunctive search space thanks to their correspondences; minimal generators and closed patterns respectively (see [6] for a study). Indeed, these particular patterns are structurally localized within the associated lattice what gives them more semantics, contrary to other patterns numerically retained (like non-derivable patterns) independently from their localization.

The rest of the paper is arranged as follows. The next section recalls the key notions used throughout this paper. Section 3 describes the concise representation based on frequent essential patterns. The disjunctive closure operator as well as its main properties are detailed in Section 4, where a new disjunctive closure-based concise representation is also introduced. The empirical evidences about the utility of our approach are provided in Section 5. Section 6 discusses the main related work.

2 Key notions

In this section, we briefly sketch the key notions used in the remainder of this paper.

Definition 1. (EXTRACTION CONTEXT) *An extraction context is a triplet $\mathcal{K}$ = ($\mathcal{O}$, $\mathcal{I}$, $\mathcal{R}$), where $\mathcal{O}$ represents a finite set of objects, $\mathcal{I}$ is a finite set of items and $\mathcal{R}$ is a binary (incidence) relation (i.e., $\mathcal{R} \subseteq \mathcal{O} \times \mathcal{I}$). Each couple $(o, i) \in \mathcal{R}$ expresses that the object $o \in \mathcal{O}$ contains the item $i \in \mathcal{I}$.*

Example 1. *In the remainder, we will consider the extraction context depicted by Table 1 with $\mathcal{O} = \{1, 2, 3, 4, 5, 6, 7\}$ and $\mathcal{I}$ = {a, b, c, d}.*

A pattern can be characterized by three kinds of supports as sketched by the following definition.

Definition 2. *[5]* (SUPPORTS OF A PATTERN) *Let $\mathcal{K}$ = ($\mathcal{O}$, $\mathcal{I}$, $\mathcal{R}$) be an extraction context. We distinguish three kinds of supports associated to a pattern I:*

	a	b	c	d
1	×			
2	×	×		
3	×		×	
4	×			×
5	×	×	×	
6	×	×		×
7	×		×	×

Table 1. An extraction context.

- *Conjunctive support:* *Supp*(I) = $\mid \{o \in \mathcal{O} \mid (\forall\ i \in I, (o,i) \in \mathcal{R})\} \mid$
- *Disjunctive support:* *Supp*$(\vee I)$ = $\mid \{o \in \mathcal{O} \mid (\exists\ i \in I, (o,i) \in \mathcal{R})\} \mid$
- *Negative support:* *Supp*$(\neg I)$ = $\mid \{o \in \mathcal{O} \mid (\forall\ i \in I, (o,i) \notin \mathcal{R})\} \mid$

A pattern I is said to be *frequent* if *Supp*(I) is greater than or equal to a user-specified minimum support threshold, denoted *minsup*. Since frequent patterns fulfill the order ideal property [7], the supersets of infrequent items will also be infrequent. The set of items $\mathcal{I}$ (and consequently the extraction context $\mathcal{K}$) will hence be considered as only containing frequent ones. Infrequent items will thus be pruned. Please also note that *Supp*$(\vee I) \geq$ *Supp*(I).

Given the respective disjunctive supports of a pattern's subsets, we are able to derive its conjunctive support using the *inclusion-exclusion identities* [5]. Furthermore, thanks to the *De Morgan's law*, we are even able to straightforwardly derive its negative support. Lemma 1 shows these important properties.

Lemma 1. (DERIVATION OF THE CONJUNCTIVE AND NEGATIVE SUPPORTS) *Let $I \subseteq \mathcal{I}$ be an arbitrary pattern. Its conjunctive and negative supports are respectively derived as follows:*

$$\begin{aligned} \textit{Supp}(I) &= \sum_{\emptyset \subset I_1 \subseteq I} (-1)^{\mid I_1 \mid \text{ - } 1}\, \textit{Supp}(\vee I_1) && (1) \\ \textit{Supp}(\neg I) &= \mid \mathcal{O} \mid - \textit{Supp}(\vee I) && (2) \end{aligned}$$

Example 2. *Consider the extraction context of Table 1. Given the respective disjunctive supports of* bc*' subsets* [3]*, its conjunctive and negative supports are inferred as follows:*

- *Supp*(bc) = $(-1)^{|bc|-1}$ *Supp*($\vee$bc) + $(-1)^{|b|-1}$ *Supp*($\vee$b) + $(-1)^{|c|-1}$ *Supp*($\vee$c) = - *Supp*($\vee$bc) + *Supp*($\vee$b) + *Supp*($\vee$c) = **- 5** + **3** + **3** = **1**.
- *Supp*($\neg$bc) = $|\mathcal{O}|$ - *Supp*($\vee$bc) = **7** - *Supp*($\vee$bc) = **7** - **5** = **2**.

3 Frequent essential pattern-based concise representation

The next definition presents the frequent essential patterns. These patterns constitute the core of the concise representation which motivates ours (*cf.* Section 1).

Definition 3. *[4]* (FREQUENT ESSENTIAL PATTERN) *Let $\mathcal{K}$ = $(\mathcal{O}, \mathcal{I}, \mathcal{R})$ be an extraction context and $I \subseteq \mathcal{I}$. I is an essential pattern* iff *Supp*$(\vee I) \neq \max\{$*Supp*$(\vee I \backslash i) \mid i \in I\}$*. An essential pattern I is also frequent if Supp$(I) \geq$ minsup.*

Example 3. *Consider the extraction context of Table 1 for minsup* = **1**. ad *is not an essential pattern since Supp*($\vee$ad) = *Supp*($\vee$a) = **7***. Whereas* bc *is an essential pattern since Supp*($\vee$bc) = **5** $\neq \max\{$*Supp*($\vee$b), *Supp*($\vee$c)$\}$ *since Supp*($\vee$b) = *Supp*($\vee$c) = **3**. bc *is also frequent since Supp*(bc) = **1** $\geq$ *minsup.*

[3] We use a separator-free form for the sets, *e.g.*, the set bc stands for {b, c}.

The set of frequent essential patterns, denoted $\mathcal{FEP}_{\mathcal{K}}$, was proven in [4] to be an order ideal in ($\mathbf{2}^{\mathcal{I}}$, $\subseteq$). The following theorem presents the frequent essential pattern-based concise representation. $\mathcal{BD}^{+}(\mathcal{FP}_{\mathcal{K}})$ denotes the set of maximal frequent patterns, which is used to detect the frequency status of an arbitrary pattern.

Theorem 1. *[4] The set $\mathcal{FEP}_{\mathcal{K}}$ of frequent essential patterns increased by $\mathcal{BD}^{+}(\mathcal{FP}_{\mathcal{K}})$ constitutes an exact concise representation of the set of frequent patterns.*

It is worth noting that in [8], this representation was shown not to be perfect, contrary to the authors' claim.

4 New disjunctive closure-based concise representation

Here we detail the main constructs related to the disjunctive closure operator [8], which will make possible to map several essential patterns into a unique element within the disjunctive search space. This is the starting point of our new concise representation.

4.1 The disjunctive closure operator

Let us start by defining the disjunctive closure operator.

Definition 4. (DISJUNCTIVE CLOSURE OPERATOR) *Let $\mathcal{K}$ = ($\mathcal{O}$, $\mathcal{I}$, $\mathcal{R}$) be an extraction context. The disjunctive closure operator h is defined as follows:*

$$h : \mathcal{P}(\mathcal{I}) \rightarrow \mathcal{P}(\mathcal{I})$$
$$I \mapsto h(I) = \{i \in \mathcal{I} \mid (\forall\, o \in \mathcal{O})\ ((o, i) \in \mathcal{R}) \Rightarrow (\exists\, i_1 \in I)((o, i_1) \in \mathcal{R})\}.$$

Roughly speaking, the disjunctive closure $h(I)$ of a pattern I is equal to the maximal set of items which *only* appear in the objects that contain at least an item of I.

Example 4. *Given the extraction context depicted by Table 1, the pattern* `bc` *is a disjunctive closed pattern since it is equal to the maximal set of items only contained in the set of objects where* `b` or `c` *appears, i.e.,* $\{\mathbf{2}, \mathbf{3}, \mathbf{5}, \mathbf{6}, \mathbf{7}\}$. *Hence,* h(`bc`) = `bc`. *While* `acd` *is not a disjunctive closed pattern since* `b` *only appears in the set of objects where at least an item of* `acd` *appears. In fact,* h(`acd`) = `abcd`.

Actually, Definition 4 gives an explicit expression of the disjunctive closure operator, free from the connection operators linking $\mathcal{P}(\mathcal{I})$ and $\mathcal{P}(\mathcal{O})$. This definition structurally characterizes the disjunctive closure of any pattern X and, hence, allows to straightforwardly compute it from any extraction context. To the best of our knowledge, our work is the first one allowing the extraction of a concise representation of frequent patterns based on a disjunctive closure operator, and, hence exploring the disjunctive search space. We will denote by $\mathcal{DCP}_{\mathcal{K}}$ the set of disjunctive closed patterns extracted from a context $\mathcal{K}$. Thanks to the closure operator h, the disjunctive search space is partitioned into distinct disjunctive equivalence classes. In the latter classes, disjunctive closed (*resp.* essential) patterns are the largest (*resp.* minimal) elements, *w.r.t.* set inclusion.

The following propositions allow to establish the relation between the smallest disjunctive closed pattern containing a pattern I and $h(I)$.

Proposition 1. *Let $I \subseteq \mathcal{I}$. $h(I)$ is the smallest disjunctive closed pattern containing I:*

$$h(I) = min_{\subseteq}\{f \in \mathcal{DCP}_{\mathcal{K}} \mid I \subseteq f\}.$$

Proposition 2. *Let $I \subseteq \mathcal{I}$. Supp($\vee$ I) = Supp($\vee$ $h(I)$).*

Proposition 3 makes possible to deduce the disjunctive closure of a pattern using the disjunctive closure of one of its subsets, while Proposition 4 establishes the link between disjunctive closed patterns and frequent essential patterns.

Proposition 3. *Let $X \subseteq \mathcal{I}$ and $Y \subseteq \mathcal{I}$ be two patterns. We then have:*

$$(X \subseteq Y \subseteq h(X)) \Rightarrow (h(Y) = h(X)).$$

Proposition 4. *Let $I \subseteq \mathcal{I}$ and $\mathcal{FP}_{\mathcal{K}}$ be the set of frequent patterns. We then have:*

$$(I \in \mathcal{FP}_{\mathcal{K}}) \Rightarrow (\exists\ f \in \mathcal{DCP}_{\mathcal{K}} \text{ and } I_1 \in \mathcal{FEP}_{\mathcal{K}} \text{ s.t. } h(I_1) = h(I) = f \text{ and } I_1 \subseteq I).$$

Proof. (*Sketch*) *The proof is based on the fact that the set $\mathcal{FEP}_{\mathcal{K}}$ is an order ideal in $(\mathbf{2}^{\mathcal{I}}, \subseteq)$ whose elements are the minimal ones in their associated disjunctive equivalence classes.*

In the remainder of the paper, we will denote by $\mathcal{EDCP}_{\mathcal{K}}$ ($\mathcal{E}$ssential $\mathcal{D}$isjunctive $\mathcal{C}$losed $\mathcal{P}$atterns) the subset of $\mathcal{DCP}_{\mathcal{K}}$ whose elements have at least a frequent essential pattern as generator. Thanks to Proposition 4, it is easy to show that the disjunctive closures of the patterns belonging to $\mathcal{BD}^{+}(\mathcal{FP}_{\mathcal{K}})$ are contained in $\mathcal{EDCP}_{\mathcal{K}}$.

Example 5. *Consider the context of Table 1. Within the disjunctive lattice sketched by Figure 1, different sets of patterns are indicated. The essential patterns are shown with bold letters, while the disjunctive closed patterns are underlined. The set $\mathcal{FEP}_{\mathcal{K}}$ induces an order ideal structure, as shown in Figure 1 for minsup = $\mathbf{1}$. Let $\mathcal{BD}^{-}(\mathcal{FEP}_{\mathcal{K}})$ be the negative border of $\mathcal{FEP}_{\mathcal{K}}$ equal to $\min_{\subseteq}\{I \in \mathcal{P}(\mathcal{I}) \setminus \mathcal{FEP}_{\mathcal{K}}\}$. The elements belonging to this border are in italic. An example of a disjunctive equivalence class, induced by the disjunctive closure operator, is also sketched. Its minimal element is the essential pattern* a *and its largest one is the disjunctive closed pattern* abcd. *Please note that if, for example, a pattern is in bold letters and is also underlined, then this means that it is both an essential pattern and a disjunctive closed one. As an indication, the patterns belonging to $\mathcal{BD}^{+}(\mathcal{FP}_{\mathcal{K}})$ are encircled.*

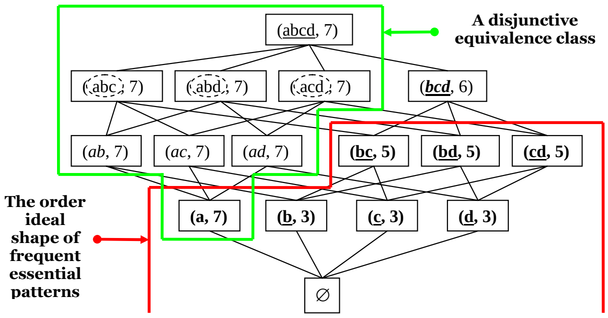

Fig. 1. The associated disjunctive lattice where each node contains a disjunctive pattern with its disjunctive support.

4.2 New disjunctive closure-based concise representation

It is commonly known that the definition of a concise representation is closely related to the way the whole set of frequent patterns will be generated starting from its elements. Suppose we have at hand the set $\mathcal{EDCP}_{\mathcal{K}}$ where each element is provided with its disjunctive support (as it is the case in [8]). We need to analyze the "tools" that will be of help in such a regeneration process. To the best of our knowledge, only the formula shown in Lemma 1 makes the link between the disjunctive support of a pattern and its conjunctive one. This formula requires knowing beforehand the disjunctive supports of the subsets of a given candidate to be able to compute its conjunctive support. Hence, an APRIORI-like regeneration is naturally advocated. This manner of regeneration consists in finding the conjunctive supports of $\mathbf{1}$-patterns, $\mathbf{2}$-patterns, and so on.

Let X be a pattern to which we are interested in retrieving its conjunctive support. Reaching X is conditioned by the fact that all its subsets (and more precisely, the immediate ones) are proven to be frequent. Indeed, the set of frequent patterns is an order ideal [7]. Hence, if a subset of X is infrequent, then X will necessarily be infrequent. Assume now that all subsets of X are frequent. At this step, the main information we have about each subset consists in its disjunctive closure (*cf.* Proposition 1) and, consequently, its different supports (*cf.* Lemma 1). If X is included in the closure of one of its immediate subsets, then we have its disjunctive closure and, hence, its disjunctive support (*cf.* Proposition 3). We can thus compute its conjunctive support. Please note that in this case, X is obviously not an essential pattern. If X is included in none of its subsets' closures, then X is necessarily an essential pattern. However, the closure of X is required to correctly compute its conjunctive support and then deduce if X is frequent or not. Nevertheless, how can we ensure that such a closure belongs to $\mathcal{EDCP}_{\mathcal{K}}$? Indeed, X can be an *infrequent* pattern and, at the same time, the *unique* generator of its disjunctive equivalence class. Hence, its closure will necessarily not be in $\mathcal{EDCP}_{\mathcal{K}}$ [(4)]. This important part was missed in [8], what motivates a careful scrutiny to correct the representation and make it really exact.

At this step of the treatment, to correctly regenerate the whole set of frequent patterns, it is clear that we need the disjunctive closures of frequent patterns (*i.e.*, $\mathcal{EDCP}_{\mathcal{K}}$), augmented by the closures *uniquely* generated by essential patterns belonging to the negative border of $\mathcal{FEP}_{\mathcal{K}}$. These latter closures do not belong to $\mathcal{EDCP}_{\mathcal{K}}$, but they bring key information when an infrequent essential pattern is reached. They are also necessarily sufficient because once an infrequent pattern is discovered all its supersets will not be treated. Hence, an important result is that $\mathcal{EDCP}_{\mathcal{K}}$ is not sufficient to ensure the exact regeneration of the whole frequent pattern set, what makes the claim of the authors in [8] incorrect. As characterized in the remainder, some closures should then be added to ensure that some candidates will not be erroneously considered as frequent whereas they are actually infrequent. These closures will form the set $\mathcal{ADCP}_{\mathcal{K}}$ ($\mathcal{A}$dded $\mathcal{D}$isjunctive $\mathcal{C}$losed $\mathcal{P}$atterns). An interesting question will be: how can we reduce the cardinality of $\mathcal{ADCP}_{\mathcal{K}}$ without affecting the exact regeneration of the whole frequent pattern set?

[4] If X is not the unique essential pattern of its disjunctive equivalence class $\mathcal{C}$, then its closure can belong to $\mathcal{EDCP}_{\mathcal{K}}$ if $\mathcal{C}$ contains at least a frequent pattern.

Let X be an infrequent essential pattern belonging to $\mathcal{BD}^-(\mathcal{FEP}_\mathcal{K})$. Let us have a look at the formula establishing the link between the conjunctive and disjunctive supports:

$$Supp(X) = \sum_{\emptyset \subset X' \subseteq X} (-1)^{|X'|-1} Supp(\vee X') = (-1)^{|X|-1} Supp(\vee X) + \sum_{\emptyset \subset X' \subset X} (-1)^{|X'|-1} Supp(\vee X').$$

Suppose that $|X|$ is even. Hence, $(-1)^{|X|-1} = \mathbf{-1}$. Assume now that we did not compute the disjunctive closure f of X. Then, two cases can arise: either X is covered by at least an element in $\mathcal{EDCP}_\mathcal{K}$ or is not covered at all (*i.e.*, $\forall f' \in \mathcal{EDCP}_\mathcal{K}$, $X \nsubseteq f'$). In the latter case, it is obvious that X is infrequent (*cf.* Proposition 4). Let us analyze the former case. Let f_1 be the smallest closure in $\mathcal{EDCP}_\mathcal{K}$ covering X. It is clear that $f \subseteq f_1$ (otherwise, the closure of X will never be f) [(5)]. Hence, $Supp(\vee f_1) \geq Supp(\vee f) = Supp(\vee X)$. Hence, if we use $Supp(\vee f_1)$ in the formula instead of $Supp(\vee X)$, the support value we obtain will be lower than or equal to the exact support of X [(6)]. This does not affect the final decision about the frequency status of X since it is infrequent and the possible decrease of its support will maintain its infrequency status. Hence, if X is an infrequent pattern of even size belonging to $\mathcal{BD}^-(\mathcal{FEP}_\mathcal{K})$, we need not compute its disjunctive closure, what consists in a very interesting pruning.

Example 6. *Consider the extraction context depicted by Table 1 for minsup =* $\mathbf{2}$. *Applying an extraction process, we obtain* $\mathcal{EDCP}_\mathcal{K} = \{(\mathsf{b}, 3), (\mathsf{c}, 3), (\mathsf{d}, 3), (\mathsf{abcd}, 7)\}$, *where each couple represents a disjunctive closed pattern and its disjunctive support. Let us regenerate the set of frequent patterns. We begin by* $\mathbf{1}$*-patterns, i.e.,* a, b, c *and* d. *The smallest closure containing* a *is* abcd. *Hence, its disjunctive support is equal to* $\mathbf{7}$, *which also corresponds to its conjunctive support. It is the same for the remaining* $\mathbf{1}$*-patterns. Thus, we find that their associated conjunctive supports are respectively equal to* $\mathbf{7}$, $\mathbf{3}$, $\mathbf{3}$ *and* $\mathbf{3}$. *We hence have the four candidates frequent. We then handle candidate* $\mathbf{2}$*-patterns. Consider the case of* bc *whose subsets are proven to be frequent. The smallest closure in* $\mathcal{EDCP}_\mathcal{K}$ *containing* bc *is* abcd. *However,* abcd *is not the actual closure of* bc. *Nevertheless, this does not affect the final decision about the frequency status of* bc. *Indeed, three cases should be distinguished:* (*i*) *if* bc *was frequent, hence its closure must belong to* $\mathcal{EDCP}_\mathcal{K}$, (*ii*) *if* bc *is not covered by the elements of* $\mathcal{EDCP}_\mathcal{K}$ *then* bc *is necessarily infrequent, otherwise,* (*iii*) *since* $|\mathsf{bc}| = \mathbf{2}$, *then* $(\text{-}1)^{|\mathsf{bc}|-1} = \mathbf{-1}$ *and hence taking a largest closure (i.e.,* abcd*), instead of the actual one (i.e.,* bc*) will decrease the result obtained thanks to Formula (1) (cf. Lemma 1), and, hence,* bc *will always be considered as infrequent and no status change can arise. Thus, the closure of* bc *is not required in the representation when* bc *is infrequent. Note that the application of Formula (1) is required independently from the frequency status of* bc *since we cannot guess its status beforehand only if it contains an infrequent subset what is not the case here.*

Unfortunately, such a pruning cannot be applied when X is of odd size. Indeed, in this case, $(-1)^{|X|-1} = \mathbf{+1}$. Thus, using $Supp(\vee f_1)$ instead of $Supp(\vee X)$ will

[5] f can be equal to f_1 if it also has a frequent essential pattern as generator.

[6] The computation of the conjunctive support of X is inevitable since we cannot beforehand predict whether it is frequent or not.

probably lead to the increase of *Supp*(X). Consequently, if X is infrequent and we augment its conjunctive support, then this may lead to a support value greater than or equal to *minsup* what clearly falsifies its frequency status. In this situation, we can further reduce the cardinality of $\mathcal{ADCP}_{\mathcal{K}}$ by only maintaining the closure f of X if it is included in at least an element of $\mathcal{EDCP}_{\mathcal{K}}$. Indeed, a pattern X is eligible to be frequent only if it is covered by a pattern of $\mathcal{EDCP}_{\mathcal{K}}$ (*cf.* Proposition 4). This can simply be done once f is computed by set inclusion operations with maximal elements of $\mathcal{EDCP}_{\mathcal{K}}$.

Example 7. *Now consider the context of Table 1 for minsup =* $\mathbf{1}$. $\mathcal{EDCP}_{\mathcal{K}}$ = $\{(\mathsf{b}, 3), (\mathsf{c}, 3), (\mathsf{d}, 3), (\mathsf{bc}, 5), (\mathsf{bd}, 5), (\mathsf{cd}, 5), (\mathsf{abcd}, 7)\}$. *As in the previous example, we begin by* $\mathbf{1}$*-patterns, i.e.,* a*,* b*,* c *and* d*. We find that their associated conjunctive supports are respectively equal to* $\mathbf{7}$*,* $\mathbf{3}$*,* $\mathbf{3}$ *and* $\mathbf{3}$*. We then treat candidate* $\mathbf{2}$*-patterns and we find that the different candidates are frequent. We now reach candidate* $\mathbf{3}$*-patterns. The unique candidate is* bcd *since all its subsets are proven to be frequent.* bcd *hence fulfills the order ideal property of frequent patterns. It is also not contained in the closure of its subsets* (*cf. Figure 1*)*.* bcd *is hence an essential pattern. If we will apply the same regeneration process to* bcd*,* abcd *will be considered as the disjunctive closure of* bcd *since it is the smallest one in* $\mathcal{EDCP}_{\mathcal{K}}$ *containing it. The conjunctive support of* bcd *will then be equal to* $\mathbf{1}$*. However, this is not true because* abcd *is not the actual disjunctive closure of* bcd*. The latter should be equal to* bcd*. Since* $|\mathsf{bcd}|$ = $\mathbf{3}$, (*-1*)$^{|\mathsf{bcd}|-1}$ = $\mathbf{+1}$ *and hence taking a largest closure* (*i.e.,* abcd)*, instead of the actual one* (*i.e.,* bcd)*, will augment the conjunctive support of* bcd*, actually equal to* $\mathbf{0}$*, which shifts its status from infrequent to frequent. The problem arises because* $\mathcal{EDCP}_{\mathcal{K}}$ *only contains closures having at least a frequent essential pattern as generator. This is not the case of* $h(\mathsf{bcd})$ *equal to* bcd *whose unique generator is obviously* bcd*. Such a closure necessarily does not belong to* $\mathcal{EDCP}_{\mathcal{K}}$ *since* bcd *is infrequent* (*its conjunctive support is equal to* $\mathbf{0}$)*. Hence, its closure must be added to the representation to ensure not including* bcd *with the set of frequent patterns during the regeneration process.*

We now give the formal definition of the set $\mathcal{ADCP}_{\mathcal{K}}$ that ensures the new representation being exact.

Definition 5. *Let* $\mathcal{EP}_{\mathcal{K}}$ *be the set of the essential patterns that can be extracted from a context* $\mathcal{K}$*. The set* $\mathcal{ADCP}_{\mathcal{K}}$ *is defined as follows:* $\mathcal{ADCP}_{\mathcal{K}} = \{h(X) \mid (X \in \mathcal{BD}^{-}(\mathcal{FEP}_{\mathcal{K}})\bigcap\mathcal{EP}_{\mathcal{K}}) \wedge ((-1)^{|X|} = \text{-}1) \wedge (\forall X' \subseteq \mathcal{I}, h(X') = h(X) \Rightarrow$ *Supp*$(X') <$ *minsup*$) \wedge (\exists f \in \mathcal{EDCP}_{\mathcal{K}}$ *s.t.* $h(X) \subset f)\}$.

To summarize, $\mathcal{ADCP}_{\mathcal{K}}$ contains closures generated by infrequent essential patterns of odd sizes belonging to $\mathcal{BD}^{-}(\mathcal{FEP}_{\mathcal{K}})$. These closures have all their corresponding essential patterns infrequent and are covered by at least one element of $\mathcal{EDCP}_{\mathcal{K}}$. It is important to mention that in $\mathcal{ADCP}_{\mathcal{K}}$, we did not consider the disjunctive closures of *infrequent non*-essential patterns belonging to $\mathcal{BD}^{-}(\mathcal{FEP}_{\mathcal{K}})$ since they are already included in $\mathcal{EDCP}_{\mathcal{K}}$ (*cf.* Proposition 3).

The concise representation $\mathcal{EDCP}_{\mathcal{K}} \bigcup \mathcal{ADCP}_{\mathcal{K}}$ will be denoted $\mathcal{DCP}_{\mathcal{K}}\_rep$.

Theorem 2. $\mathcal{DCP}_{\mathcal{K}}\_rep$ *is an exact concise representation of* $\mathcal{FP}_{\mathcal{K}}$*.*

The proof of Theorem 2 can be treated as a naive algorithm for deriving frequent patterns and their associated supports.

In addition to the exact retrieval of frequent patterns as well as their various supports, $\mathcal{DCP}_{\mathcal{K}}\_rep$ presents three other main properties:
1. Homogeneity: $\mathcal{DCP}_{\mathcal{K}}\_rep$ only involves disjunctive closed patterns (*vs.* $\mathcal{FEP}_{\mathcal{K}} \bigcup \mathcal{BD}^{+}(\mathcal{FP}_{\mathcal{K}})$). Hence, it ensures the homogeneity of the representation since all its elements are provided with the same kind of support; the disjunctive one. They also have the same structural properties. Indeed, they are the top elements of their associated equivalence classes within the disjunctive search space.
2. Small size: In [8], the size of $\mathcal{EDCP}_{\mathcal{K}}$ is shown to be significantly smaller than those of the best known concise representations. In addition, the size of $\mathcal{ADCP}_{\mathcal{K}}$ is very small since its elements must fulfill many easy-to-check constraints. Hence, the size of $\mathcal{DCP}_{\mathcal{K}}\_rep$ will be, in most cases, smaller than those of the other representations.
3. Low regeneration cost: It is worth mentioning that our concise representation allows retrieving the conjunctive support faster than from (closed) non-derivable patterns [2, 3]. Indeed, for a pattern X *s.t.* $|X| = n$, the retrieval process of *Supp*(X) from these representations requires the costly evaluation of $\mathbf{2}^{n}$ deduction rules based on Bonferroni-inequalities [9]. The computation cost for inferring supports is then awfully high. While the retrieval of *Supp*(X) from our concise representation only needs to evaluate a unique inclusion-exclusion identity. Furthermore, it allows the straightforward retrieval of the disjunctive and negative supports of frequent patterns.

5 Experimental results

We compare, through various experiments, the size of our concise representation to those of the exact ones based on frequent closed, (closed) non-derivable and essential patterns. This is done in the most critical cases, *i.e.*, for strongly correlated datasets [7]. Indeed, within such datasets, the ratio between the cardinality of the frequent pattern set and those of concise representations is high. Thus, we are in the most interesting cases. Moreover, equivalence classes extracted from sparse datasets are often reduced to the associated generators and cannot be compacted anymore. This makes the size reduction rates brought by concise representations meaningless in such datasets. Due to lack of space, we only summarize the main results in this section.

All experiments were carried out on a PC equipped with a 1.73GHz Centrino Duo Core and 2GB of main memory, and running the Linux version Fedora Core 6 (with 2GB of swap memory). Results are shown in Table 5. The abbreviation "$\mathcal{FP}_{\mathcal{K}}\_set$"(*resp.* "$\mathcal{FCP}_{\mathcal{K}}\_rep$" [8], "$\mathcal{NDP}_{\mathcal{K}}\_rep$", "$\mathcal{CNDP}_{\mathcal{K}}\_rep$", and "$\mathcal{FEP}_{\mathcal{K}}\_rep$") is used to stand for the set of frequent patterns (*resp.* frequent closed, non-derivable, closed non-derivable and essential pattern-based concise representation). It is important to note that in the experimental results given in [3], the authors have chosen a specific interval of *minsup* for each dataset to extract $\mathcal{CNDP}_{\mathcal{K}}\_rep$. However, we noticed that their program abruptly comes to an end with an execution error beyond these intervals. Therefore, we use the symbol "-" to designate a case where an execution error occurred. At a glance, we can also deduce the following assertions:
1. Necessity to set up concise representations: Indeed, their respective sizes *w.r.t.* that

[7] These datasets are available at: *http://fimi.cs.helsinki.fi/data.*

[8] Source codes for extracting frequent (closed) patterns are available at: *http://fimi.cs.helsinki.fi/src.*

of the set of frequent patterns clearly show their utility and potential benefits. In particular, even for high *minsup* values, the cardinality of the introduced concise representation is considerably reduced.

2. Effectiveness of the proposed concise representation: Indeed, for CHESS, CONNECT and PUMSB datasets, the size of $\mathcal{DCP}_{\mathcal{K}}\_rep$ is significantly reduced compared to those of the remaining concise representations, while offering different kinds of patterns' supports.

3. Scalability of $\mathcal{DCP}_{\mathcal{K}}\_rep$: It is easily observable that, in most cases, the cardinality of $\mathcal{DCP}_{\mathcal{K}}\_rep$ is less sensible to the variation of *minsup* than those of the other concise representations.

4. Absence of an outstanding concise representation: For example, in some cases, the size of $\mathcal{DCP}_{\mathcal{K}}\_rep$ is slightly greater than the size of the other concise representations (*e.g.*, MUSHROOM for *minsup* = 5%).

***minsup* (%)**	$\vert\mathcal{FP}_{\mathcal{K}}\_set\vert$	$\vert\mathcal{FCP}_{\mathcal{K}}\_rep\vert$	$\vert\mathcal{NDP}_{\mathcal{K}}\_rep\vert$	$\vert\mathcal{CNDP}_{\mathcal{K}}\_rep\vert$	$\vert\mathcal{FEP}_{\mathcal{K}}\_rep\vert$	$\vert\mathcal{DCP}_{\mathcal{K}}\mathbf{\_rep}\vert$
CONNECT						
90	27, 127	3, 486	199	177	398	$\mathbf{22}$
70	4, 129, 839	35, 875	545	491	1, 710	$\mathbf{161}$
50	88, 324, 400	130, 112	1, 397	-	5, 063	$\mathbf{589}$
30	1, 331, 673, 367	460, 356	3, 221	-	14, 083	$\mathbf{1,\ 986}$
MUSHROOM						
40	565	140	146	117	151	$\mathbf{91}$
20	53, 583	1, 197	1, 143	731	1, 258	$\mathbf{941}$
10	574, 431	4, 885	4, 347	2, 655	6, 530	$\mathbf{5,\ 457}$
5	3, 755, 511	12, 843	11, 569	6, 546	24, 407	$\mathbf{20,\ 554}$
CHESS						
90	622	498	95	93	118	$\mathbf{43}$
70	48, 731	23, 892	684	669	1, 482	$\mathbf{420}$
50	1, 272, 932	369, 450	3, 425	3, 341	14, 272	$\mathbf{1,\ 971}$
30	37, 282, 962	5, 316, 467	15, 147	-	147, 777	$\mathbf{8,\ 824}$
PUMSB						
90	2, 607	1, 467	586	460	788	$\mathbf{318}$
80	142, 156	33, 308	3, 642	2, 136	6, 251	$\mathbf{1,\ 079}$
70	2, 698, 654	241, 259	7, 875	4, 564	18, 318	$\mathbf{2,\ 143}$
60	19, 529, 991	1, 074, 627	21, 323	-	54, 644	$\mathbf{5,\ 550}$
50	165, 903, 540	7, 121, 264	47, 764	-	232, 581	$\mathbf{11,\ 551}$

Table 2. Size of the different concise representations for benchmark datasets.

6 Discussion

First of all, let us make an alignment between the disjunctive search space and the conjunctive one. We will hence find that an essential pattern is the mapping of the concept of *minimal generator* (*aka key pattern* and *free-set* in the literature, see [6] for references) when the conjunctive search space is considered. While the disjunctive closed patterns are the mapping of conjunctive ones [1].

The concepts of essential and disjunctive closed patterns can be considered as particular cases of *composite items* [10] where the disjunction of items is used to compose new items, the composite ones. This is an attempt towards making useful infrequent items in some applications. For example, consider the context of Table 1 and let *minsup* = $\mathbf{4}$, b and c are hence infrequent items since their support is equal to $\mathbf{3}$. Nevertheless, the support of $b\vee c$ is equal to $\mathbf{5}$ and, hence, *Supp*($b\vee c$) $\geq$ *minsup*. $b\vee c$ will be considered as a new item (a composite one) even if, actually it is composed of two items. It will be used during the mining process since it is frequent what makes b and c useful.

It is important to make the link between our work and that of Zhao *et al*. Indeed, in [11], the authors proposed connection operators to link $\mathcal{P}(\mathcal{I})$ and $\mathcal{P}(\mathcal{O})$ for the case of disjunctive Boolean expressions. Nevertheless, their definition of the operator linking $\mathcal{P}(\mathcal{O})$ to $\mathcal{P}(\mathcal{I})$ depends on that ensuring the opposite direction and was not independently given from any other operator. Furthermore, they neither proposed the expression of the resulting closure operator nor carried out a thorough analysis of inherent structural properties.

The disjunction operator (*i.e.*, the operator $\vee$) has also been used to define some concise representations only exploring the conjunctive search space, like those based on disjunction-free sets and (generalized) disjunction-free generators [12] [(9)]. This required the introduction of what is called *disjunctive rule*. Such a rule has a premise part composed by a conjunction of items and a conclusion part, distinct from the premise one, containing a specified number of items linked using the disjunction operator [12].

Some works [13, 14] were interested in using disjunction within association rules to define what is called generalized association rules. These rules grasped the interest of many researchers since they offer wealthier types of knowledge in many applications. In addition to the inclusive disjunction operator, *i.e.*, the operator $\vee$, the authors in [13] were also interested in the exclusive disjunction operator, denoted $\oplus$. In [14], the author mainly focusses on association rules having conclusions containing mutually exclusive items, *i.e.*, the presence of one of them leads to the absence of the others, what is expressed in [13] using the operator $\oplus$. Other forms of generalized association rules were also described in [15].

7 Conclusion and future work

In this paper, we presented a new disjunctive closure operator as well as its main properties. Based on this operator, we introduced a new concise representation which corrects the claim of [8] where the associated representation can miss some cases. This required the addition of few further elements what ensures the correctness of the whole regeneration process of frequent patterns. In addition to interesting compactness rates, our concise representation allows a straightforward computation of the disjunctive and negative supports. The experimental results showed that, in most cases, its size is significantly smaller than those of the best known concise representations. It is worth noting that our approach can easily be extended when negative items are handled.

Other avenues for future work mainly address the following points: First, due to space limitations here, we intend to address as next step the complexity time issue (generation and derivability) of our representation *vs.* those of the literature. In this respect, other algorithms for mining conjunctive closed patterns could be adapted to disjunctive ones, both breadth-first search algorithms and depth-first ones. Second, a structural characterization of disjunctive closed patterns *w.r.t.* existing frameworks like the k-free sets [12] will be done. Another important task consists in overcoming the lack in the literature of semantics' studies related to concise representations. The study of the possible extension of our representation to other pattern classes should also be examined. Finally, the extraction of generalized association rules will be thoroughly

[9] We did not use these representations in our experiments since $\mathcal{NDP}_{\mathcal{K}}\_rep$ (and consequently, $\mathcal{CNDP}_{\mathcal{K}}\_rep$) is shown in [2] to provide better results.

addressed. Indeed, setting up a theoretical framework that includes different kinds of operators is of paramount importance for jumping beyond standard association rules.

Acknowledgements: We thank the anonymous reviewers for their helpful comments. We also would like to express our thanks to Toon Calders and Bart Goethals for providing the source code allowing the extraction of frequent non-derivable patterns, Alain Casali and Stéphane Lopes for the source code allowing the extraction of the concise representation based on frequent essential patterns, and, Juho Muhonen and Hannu Toivonen for the source code allowing the extraction of frequent closed non-derivable patterns. This work is partly supported by the French-Tunisian CMCU project 05G1412.

References

1. Pasquier, N., Bastide, Y., Taouil, R., Stumme, G., Lakhal, L.: Generating a condensed representation for association rules. Journal of Intelligent Information Systems, Kluwer Academic Publisher, **volume 24(1)** (2005) 25–60
2. Calders, T., Goethals, B.: Non-derivable itemset mining. Data Mining and Knowledge Discovery (DMKD), Springer, **volume 14(1)** (2007) 171–206
3. Muhonen, J., Toivonen, H.: Closed non-derivable itemsets. In: Proceedings of the 10th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD 2006), Springer-Verlag, LNAI, volume 4212, Berlin, Germany. (2006) 601–608
4. Casali, A., Cicchetti, R., Lakhal, L.: Essential patterns: A perfect cover of frequent patterns. In: Proceedings of the 7th International Conference on Data Warehousing and Knowledge Discovery (DaWaK 2005), Springer-Verlag, LNCS, volume 3589, Copenhagen, Denmark. (2005) 428–437
5. Galambos, J., Simonelli, I.: Bonferroni-type inequalities with applications. Springer (2000)
6. Hamrouni, T., Ben Yahia, S., Mephu Nguifo, E.: Succinct minimal generators: Theoretical foundations and applications. To appear in the International Journal of Foundations of Computer Science (IJFCS). (2007)
7. Ganter, B., Wille, R.: Formal Concept Analysis. Springer (1999)
8. Hamrouni, T., Ben Yahia, S., Mephu Nguifo, E.: A new exact concise representation based on disjunctive closure. In: Proceedings of the 2nd Jordanian International Conference on Computer Science and Engineering (JICCSE 2006), Al-Balqa, Jordan. (2006) 361–373
9. Mielikäinen, T., Panov, P., Dzeroski, S.: Itemset support queries using frequent itemsets and their condensed representations. In: Proceedings of the 9th International Conference Discovery Science (DS 2006), Springer-Verlag, LNCS, volume 4265, Barcelona, Spain. (2006) 161–172
10. Ye, X., Keane, J.A.: Mining composite items in association rules. In: Proceedings of the 1997 IEEE International Conference on Systems, Man, and Cybernetics (SMC 1997), Hyatt Orlando, Orlando, Florida, USA. (1997) 1367–1372
11. Zhao, L., Zaki, M.J., Ramakrishnan, N.: BLOSOM: A framework for mining arbitrary Boolean expression. In: Proceedings of the 12th International Conference on Knowledge Discovery and Data Mining (KDD 2006), Philadelphia, PA, USA. (2006) 827–832
12. Calders, T., Rigotti, C., Boulicaut, J.F.: A survey on condensed representations for frequent sets. In: Constraint Based Mining and Inductive Databases, Springer-Verlag, LNAI, volume 3848. (2005) 64–80
13. Nanavati, A.A., Chitrapura, K.P., Joshi, S., Krishnapuram, R.: Mining generalised disjunctive association rules. In: Proceedings of the 10th International Conference on Information and Knowledge Management (CIKM 2001), Atlanta, Georgia, USA. (2001) 482–489
14. Kim, H.D.: Complementary occurrence and disjunctive rules for market basket analysis in data mining. In: Proceedings of the 2nd IASTED International Conference Information and Knowledge Sharing (IKS 2003), Scottsdale, AZ, USA. (2003) 155–157
15. Grün, G.A.: New forms of association rules. Technical Report TR 1998-15, School of Computing Science, Simon Fraser University, Burnaby, BC, Canada (1998)