Sweeping the disjunctive search space towards mining new exact concise representations of frequent itemsets

T. Hamrouni a,b, S. Ben Yahia a,*, E. Mephu Nguifo c

a Computer Science Department, Faculty of Sciences of Tunis, University Campus, 1060 Tunis, Tunisia
b CRIL-CNRS, Lille Nord University, Rue de l’université, 62307 Lens cedex, France
c LIMOS-CNRS, Blaise Pascal University (Clermont Ferrand 2), Campus des cézeaux, 63173 Aubière cedex, France

1. Introduction

In data mining, frequent itemsets and association rules are among the most popular research topics [25,53,59]. These pattern classes are closely related since the extraction of the first is usually considered as a starting point for that of the second. In practice, to be able to manage the overwhelming quantity of patterns that can be drawn from real-life datasets [51], many approaches advocate the extraction of concise (or condensed) representations [34]. In this respect, concise representations of frequent itemsets were used in various applications where frequent itemsets and their associated supports are useful [13,37]. Interestingly, the use of concise representations was extended to many pattern classes. For example, they are at the roots of different proposals aiming at concisely representing pattern classes such as association rules [17], associative classification rules [4], inter-transaction itemsets [29], sequential patterns [3,32,43], graphs [57], trees [2], minimal transversals [27], multidimensional patterns [16,42], etc.

* Corresponding author. Tel.: +216 71 885 190.
E-mail addresses: tarek.hamrouni@fst.rnu.tn, hamrouni@cril.univ-artois.fr (T. Hamrouni), sadok.benyahia@fst.rnu.tn (S.B. Yahia), engelbert.mephu-nguifo@univ-bpclermont.fr, mephu@isima.fr (E.M. Nguifo).

0169-023X/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.datak.2009.05.001

Please cite this article in press as: T. Hamrouni et al., Sweeping the disjunctive search space towards mining new exact concise representations of frequent itemsets, Data Knowl. Eng. (2009), doi:10.1016/j.datak.2009.05.001
Within the traditional association analysis, the major concern has been the conjunctive connector linking items [17]. Such a connector conveys information about the frequency of the simultaneous occurrence (or co-occurrence) of items in a dataset. The induced search space of itemsets is hence called the conjunctive search space. In fact, the use of the conjunctive connector was mainly motivated by the original application pertaining to market basket analysis [1]. More recently, a growing number of approaches has explored this latter space aiming at getting out an information lossless nucleus of itemsets, from which the remaining ones can be derived. Such a reduced set is better known as exact concise representation. Beyond high compactness rates, an exact concise representation makes it possible to guess the frequency status of an itemset and to exactly retrieve its exact support in the case of that itemset is (potentially) interesting w.r.t. statistical metrics. Many concise representations were proposed in the literature, like those based on frequent closed itemsets [40], minimal generators [31], disjunction-free sets [10,11], (generalized) disjunction-free generators [28], non-derivable itemsets [12], closed non-derivable itemsets [38], and essential itemsets [15], to cite but a few. Considering the set of frequent itemsets as data, all those representations follow the minimum description length (MDL) principle [45] which is based on the following insight: any regularity in the data can be used to describe the data using fewer symbols than the number of symbols needed to describe the data literally [22].

In practice, some situations can arise highlighting the importance of taking into account other kinds of relations between items, like mutually exclusive or complementary occurrences [54]. Suppose that a market basket data is under treatment, and the manager is searching for items $c_1, c_2, \ldots, c_n$ whose selling implies that of at least one of two competitive products $a$ and $b$ (or probably both), i.e., the items fulfilling the condition: $c_1 \lor c_2 \lor \ldots \lor c_n \Rightarrow a \lor b$ is always true and conveys information about the sold items simultaneously to $a$ or $b$. Since the disjunctive connector $\lor$ is inclusive, the simultaneous selling of $c_i$ and $c_j$ ($i \neq j$) is possible. On the other hand, in a text mining application related to text translation from a language $l_1$ to a language $l_2$, an analyst may be interested in the possible translations in the language $l_2$ of a given term $t$ belonging to the language $l_1$. In this respect, $t$ may have several translations $t_1, t_2, \ldots, t_n$ in the language $l_2$ according to its usage context. Thus, a rule like $t_1 \lor t_2 \lor \ldots \lor t_n \Rightarrow t$ is interesting since it summarizes the possible translation of $t$. In both cases, more computations may be performed to get more precise information about the effect of a given product (resp. terms) among $c_i$ (resp. $t_i$) on the appearance of $a$ and $b$ (resp. $t$). Various other applications of disjunctive itemsets are possible in the contexts of market basket analysis [39], medical data analysis [44], social network analysis and bioinformatics [61], software change impact analysis [26], feature model mining [46], etc.

In such situations, the disjunctive connector linking itemsets can bring key information as well as a summarizing method of conveyed knowledge about the complementary occurrence of items in a dataset. Thus, due to the close link between frequent itemsets and association rules, it is more advantageous to mine a concise representation of frequent itemsets that give direct access to the disjunctive support of frequent itemsets. Such a representation can be used as a starting point for mining such disjunctive rules based on frequently occurring itemsets.

To the best of our knowledge, the exact concise representation based on frequent essential itemsets is the unique representation offering this interesting feature through its exploration of the disjunctive search space. In this space, itemsets are characterized by their respective disjunctive supports. Thus, an itemset verifies an element of a dataset (or transaction) if one of its items belongs to this transaction. With respect to set inclusion, an essential itemset is the conjunctive search space. In this space, itemsets are characterized by their respective disjunctive supports. Thus, an itemset verifies an element of a dataset (or transaction) if one of its items belongs to this transaction. With respect to set inclusion, an essential itemset is the minimal set of items among those itemsets characterizing a common set of transactions. To bridge both disjunctive and conjunctive search spaces, the inclusion–exclusion identities [20] are of use to deduce the conjunctive supports of itemsets starting from their disjunctive supports. Hence, this representation offers a basis for straightforwardly deriving the conjunctive, disjunctive and negative frequencies of a pattern [14,15].

In spite of such interesting structural and compactness properties, this exact concise representation presents two major limitations:

1. It is not self-contained in the sense that the set of frequent essential itemsets does not make it possible by itself to decide whether an itemset is frequent or not. Hence, to get out this information, this set has to be burdened by the positive border of frequent itemsets, composed by the frequent maximal itemsets [6];
2. Several essential itemsets may characterize the same set of data and, therefore, they present a certain form of redundancy.

In this respect, a compelling and thriving issue is to find a closure operator related to essential itemsets in the sake of getting a more reduced concise representation, following the MDL principle. Indeed, a gain in compactness terms can be reached thanks to the non-injectivity property of the closure operator since many essential itemsets will be mapped into a single element within the disjunctive search space.

In this work, the main contribution is to introduce a new exact concise representation based on the disjunctive closure of essential itemsets. To do so, we have to introduce a new closure operator associated to the disjunctive search space and a thorough study of its theoretical properties. The targeted representation aims at palliating the limitations of that based on frequent essential itemsets as follows:

1. Getting out a more compact representation than that based on frequent essential itemsets by exploiting the non-injectivity property of the introduced closure operator. In fact, we have to only retain the disjunctive closed itemsets that ensure to exactly recovering the whole set of frequent itemsets.

Please cite this article in press as: T. Hamrouni et al., Sweeping the disjunctive search space towards mining new exact concise representations of frequent itemsets, Data Knowl. Eng. (2009), doi:10.1016/j.datak.2009.05.001
2. Ensuring the homogeneity of the obtained concise representation by only keeping itemsets characterized by their disjunctive support.

Exhaustive experiments, focusing on the compactness aspect, show the effectiveness of our concise representation compared to the pioneering representations of the literature. Here again, the MDL principle allows for an objective comparison of alternative models regardless of their form or number of parameters in case the interest is in model selection [45]. In addition, to the best of our knowledge, our work is the first one allowing the extraction of such a cover thanks to a disjunctive closure operator.

The remainder of the paper is organized as follows: Section 2 presents the background used throughout the paper. Section 3 details the disjunctive closure operator and its main properties. Then, Section 4 describes the induced structural properties of the disjunctive search space. New disjunctive closure-based representations of (frequent) itemsets are then introduced in Section 5. The empirical evidences about the utility of our approach are provided in Section 6. We also discuss related work in Section 7. The paper ends with a conclusion of our contributions and sketches forthcoming issues in Section 8.

2. Background

2.1. Basic concepts

In this section, we present the basic concepts that will be of use in the remainder.

**Definition 1.** A dataset is a triplet $D = (\mathcal{F}, \mathcal{I}, \mathcal{R})$, where $\mathcal{F}$ represents a finite set of transactions, $\mathcal{I}$ is a finite set of items and $\mathcal{R}$ is a binary (incidence) relation (i.e., $\mathcal{R} \subseteq \mathcal{F} \times \mathcal{I}$). Each couple $(t, i)$ belonging to $\mathcal{R}$ expresses that the transaction $t \in \mathcal{F}$ contains the item $i \in \mathcal{I}$.

**Example 1.** An example of a dataset $D$ is depicted by Table 1 with $\mathcal{F} = \{1, 2, 3, 4, 5, 6, 7, 8\}$ and $\mathcal{I} = \{A, B, C, D, E\}$.

The following definition presents the different types of supports that can be assigned to an itemset.

**Definition 2.** Let $D = (\mathcal{F}, \mathcal{I}, \mathcal{R})$ be a dataset. We distinguish three kinds of supports associated to an itemset $I$:

- **Conjunctive support:** $\text{Supp}(I) = \{|t \in \mathcal{F}|(\forall i \in I, (t, i) \in \mathcal{R})\}$.
- **Disjunctive support:** $\text{Supp}(\lor I) = \{|t \in \mathcal{F}|(\exists i \in I, (t, i) \in \mathcal{R})\}$, and,
- **Negative support:** $\text{Supp}(\neg I) = \{|t \in \mathcal{F}|(\forall i \in I, (t, i) \notin \mathcal{R})\}$.

Roughly speaking, the different supports are defined as follows:

- $\text{Supp}(I)$ is the number of transactions containing all items of $I$. In this case, the itemset $I$ can be seen as a conjunction of items $(i_1 \land i_2 \land \ldots \land i_n)$ such that the appearance of one of its items is conditioned by the appearance of all remaining items to say that $I$ satisfies a given transaction.
- $\text{Supp}(\lor I)$ is the number of transactions containing at least one item of $I$. In this case, the itemset $I$ can be seen as a disjunction of items $(i_1 \lor i_2 \lor \ldots \lor i_n)$ such that the presence of one item of $I$ in a given transaction is sufficient to satisfy it independently from the remaining items.
- $\text{Supp}(\neg I)$ is the number of transactions that do not contain any item of $I$.

**Example 2.** Consider the dataset in Table 1. The different supports that can be associated to the itemset $AE$ are: $\text{Supp}(AE) = 3$, $\text{Supp}(\lor AE) = 7$, $\text{Supp}(\neg AE) = 1$.

The next proposition summarizes important properties related to the itemsets supports.

---

Please cite this article in press as: T. Hamrouni et al., Sweeping the disjunctive search space towards mining new exact concise representations of frequent itemsets, Data Knowl. Eng. (2009), doi:10.1016/j.datak.2009.05.001
Proposition 1. Let $i \in \mathcal{I}$ and $I, I_1 \subseteq \mathcal{I}$. The following properties hold:

- $\text{Supp}(i) = \text{Supp}(\lor i)$,
- $\text{Supp}(I) \subseteq \text{Supp}(\lor I)$,
- If $I \subseteq I_1$, then $\text{Supp}(I) \supseteq \text{Supp}(I_1)$,
- If $I \subseteq I_1$, then $\text{Supp}(\lor I) \subseteq \text{Supp}(\lor I_1)$.

An itemset $I$ is said to be frequent if $\text{Supp}(I)$ is greater than or equal to a minimum support threshold, denoted $\text{minsupp}$. The set of frequent itemsets will be denoted $\mathcal{F}$. It was shown in [1] that the frequency is an anti-monotone property. Since the supersets of infrequent itemsets will be infrequent, the set $\mathcal{F}$ (and consequently the dataset $\mathcal{D}$) will be reduced to frequent items. Infrequent itemsets will thus be pruned. The set $\mathcal{F}$, partially ordered with set inclusion, induces an order ideal in $(2^\mathcal{I}, \subseteq)$. An order ideal splits the power-set of items into two disjoint parts: the first contains itemsets fulfilling the associated constraint. For the set $\mathcal{F}$, the constraint is “to be frequent” and it hence contains frequent itemsets. While the second part contains itemsets not fulfilling the constraint, namely infrequent itemsets in the case of the frequency constraint. Both parts are delimited thanks to a positive and a negative border, respectively [35]. These borders are presented thanks to Definition 3.

Definition 3. Let $(2^\mathcal{I}, \subseteq)$ be a partially ordered set of elements and $S$ be a subset of $2^\mathcal{I}$ s.t. $S$ is an order ideal in $(2^\mathcal{I}, \subseteq)$. $S$ can be represented by its positive border $\mathcal{Bd}^+(S)$ or its negative border $\mathcal{Bd}^-(S)$ defined as follows:

$$\mathcal{Bd}^+(S) = \max_{\mathcal{I}} \{ I \in S \},$$

$$\mathcal{Bd}^-(S) = \min_{\mathcal{I}} \{ I \in 2^\mathcal{I} \setminus S \}.$$ 

Given the respective disjunctive supports of an itemset subsets, we are able to derive its conjunctive support using the inclusion–exclusion identities [20]. 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. Let $I \subseteq \mathcal{I}$ be an arbitrary itemset. Its conjunctive and negative supports are respectively derived as follows [20]:

$$\text{Supp}(I) = \sum_{i \in \mathcal{I}} (-1)^{|I|} \text{Supp}(\lor I),$$

$$\text{Supp}^-(I) = |\mathcal{I}| - \text{Supp}(\lor I).$$

Example 3. Consider the dataset of Table 1. Given the respective disjunctive supports of AE subsets, its conjunctive and negative supports are inferred as follows:

- $\text{Supp}(\text{AE}) = (-1)^{|\text{AE}|} \text{Supp}(\lor \text{AE}) + (-1)^{|\text{AE}|-1} \text{Supp}(\lor A) + (-1)^{|A|} \text{Supp}(\lor E) = \text{Supp}(\lor A) + \text{Supp}(\lor E) = -7 + 6 = 3$,
- $\text{Supp}^-(\text{AE}) = |\mathcal{I}| - \text{Supp}(\lor \text{AE}) = 8 - \text{Supp}(\lor \text{AE}) = 8 - 7 = 1$.

2.2. Frequent essential itemset-based concise representation

The next definition presents the frequent essential itemsets [15].

Definition 4. An itemset $I \subseteq \mathcal{I}$ is essential if $\text{Supp}(\lor I) \geq \max\{\text{Supp}(\lor (I \setminus \{i\})) | i \in I\}$. $I$ is a frequent essential itemset if it is simultaneously frequent and essential.

Example 4. Consider the dataset of Table 1 for $\text{minsupp} = 2$. The itemset ABC is not an essential itemset since $\text{Supp}(\lor \text{ABC}) = \text{Supp}(\lor \text{AB}) = 7$. Whereas, AE is an essential itemset. Indeed, $\text{Supp}(\lor \text{AE}) = 7 > \max\{\text{Supp}(\lor \text{A}), \text{Supp}(\lor \text{E})\}$, since $\text{Supp}(\lor \text{A}) = 6$, while $\text{Supp}(\lor \text{E}) = 4$. The itemset AE is also frequent, since $\text{Supp}(\text{AE}) = 3 \geq \text{minsupp}$.

The set of frequent essential itemsets, denoted by $\mathcal{FEI}$, which can be drawn from a dataset $\mathcal{D}$ was proven in [15] to fulfill the anti-monotone property. However, to ensure the exact regeneration of frequent patterns, this set must be augmented by the set $\mathcal{Bd}^-(\mathcal{F})$ of maximal frequent itemsets [6].

3. The disjunctive closure operator and associated properties

3.1. Description

The basic idea of this new concise representation is to apply a closure operator on frequent essential itemsets to obtain a more compact concise representation while preserving their interesting properties. As this will be structurally characterized...
in the next section, the disjunctive itemsets will be divided into subsets, and each subset simply represented by a unique element: the disjunctive closed itemset. This relies on the non-injectivity property of any closure operator. The application of this operator makes it possible to reduce the number of itemsets to be retained in the representation while being able to regenerate the whole set of frequent itemsets without information loss.

The targeted operator is different from that applied in the case of conjunctively closed itemsets [40]. Indeed, essential itemsets are characterized within the “disjunctive search space” and no more within the “conjunctive search space”. Thus, as shown by Definition 4, they are characterized by their disjunctive supports and no more by their conjunctive supports.

Hence, a new disjunctive closure operator has to be devised.

The presentation of the new disjunctive closure requires that we define the corresponding applications ensuring the link from the power-set of items \( \mathcal{P}(\mathcal{F}) \) to that of transactions \( \mathcal{P}(\mathcal{T}) \) and vice versa.

**Definition 5.** Let \( \mathcal{D} = (\mathcal{T}, \mathcal{F}, \mathcal{A}) \) be a dataset. The operators ensuring the connection between the \( \mathcal{P}(\mathcal{F}) \) and \( \mathcal{P}(\mathcal{T}) \) are as follows:

\[
\begin{align*}
  f : \mathcal{P}(\mathcal{F}) &\rightarrow \mathcal{P}(\mathcal{F}) \\
  T \mapsto f(T) &= \{ i \in \mathcal{F} | (\exists t \in T)(((t, i) \in \mathcal{A}) \land ((\forall t_1 \in \mathcal{F} \setminus T)((t_1, i) \notin \mathcal{A})))\} \\
  g : \mathcal{P}(\mathcal{F}) &\rightarrow \mathcal{P}(\mathcal{T}) \\
  h \rightarrow g(h) &= \{ t \in \mathcal{T} | (\exists i \in \mathcal{F} \land ((t, i) \in \mathcal{A}))\}
\end{align*}
\]

Let us semantically explain these operators. With respect to set inclusion, \( f(T) \) is the maximal set of items which only appear in the transactions of \( T \). Dually, \( g(I) \) is the largest set of transactions which contain at least an item of \( I \).

**Example 5.** If we consider the dataset depicted by Table 1, we then have: \( f(\{2\}) = 0 \), \( f(\{1, 4, 6, 7\}) = \{C\} \) and \( g(\{A, B, D\}) = \{1, 2, 3, 4, 5, 7, 8\} \), \( g(\{C\}) = \{1, 3, 4, 6, 7, 8\} \).

Based on the operators introduced in **Definition 5**, we present the compound operators \( f \circ g \) and \( g \circ f \).

**Definition 6.** Let \( \mathcal{D} = (\mathcal{T}, \mathcal{F}, \mathcal{A}) \) be a dataset. Let \( f \) and \( g \) be the operators as introduced in **Definition 5**. We define the resulting compound operators as follows:

\[
\begin{align*}
  h = f \circ g : \mathcal{P}(\mathcal{F}) &\rightarrow \mathcal{P}(\mathcal{F}) \\
  I \mapsto h(I) &= \{ i \in \mathcal{F} | (\forall t \in \mathcal{T})((t, i) \in \mathcal{A}) \Rightarrow (\exists i_1 \in I)((t, i_1) \in \mathcal{A}))\} \\
  h' = g \circ f : \mathcal{P}(\mathcal{F}) &\rightarrow \mathcal{P}(\mathcal{T}) \\
  T \mapsto h'(T) &= \{ t \in \mathcal{T} | (\exists i \in \mathcal{F} \land ((t, i) \in \mathcal{A})) \Rightarrow (\forall t_1 \in \mathcal{T} \land (t_1, i) \notin \mathcal{A}))\}
\end{align*}
\]

Let us semantically explain these compound operators. Let \( I \) be an itemset, \( h(I) = f \circ g(I) \) is equal to the largest set of items which only appear in the transactions that contain at least an item of \( I \). Let \( T \) be a set of transactions, \( h'(T) = g \circ f(T) \) is equal to the set of transactions that contain at least an item only appearing in the transactions of \( T \).

**Example 6.** Consider the dataset of Table 1. We have: \( h(\{E\}) = f \circ g(\{E\}) = f(\{1, 2, 6, 8\}) = E \), \( h(\{A, D\}) = f(\{2, 3, 4, 5, 6, 7, 8\}) = ABCDE \), and \( h'(\{3\}) = g \circ f(\{3\}) = g(\{0\}) = \emptyset \), \( h'(\{1, 4, 6, 7\}) = g \circ f(\{1, 4, 6, 7\}) = g(C) = \{4, 6, 7\} \).

Using itemset supports, we can also characterize the disjunctive closure of an arbitrary itemset as shown by the following definition.

**Definition 7.** The disjunctive closure of an itemset \( I \) is equal to: \( h(I) = I \cup \{ i \in \mathcal{F} \setminus \text{Supp}(\lor I) \} \text{Supp}(\lor I) = \text{Supp}(\lor (I \cup \{i\})) \).

Thus, \( h(I) \) is the maximal itemset, w.r.t. set inclusion, containing \( I \) and having the same disjunctive support. It can be obtained incrementally if we have the disjunctive support of the proper supersets of \( I \) by considering items that do not change the disjunctive supports of \( I \). The appearance of these itemsets in the dataset is consequently dependent on that of a nonempty subset of \( I \).

**Example 7.** Consider the dataset of Table 1. Let us look for the disjunctive closure of \( AB \). We have \( \text{Supp}(\lor AB) = \text{Supp}(\lor ABC), \text{Supp}(\lor AB) \neq \text{Supp}(\lor ABD) \), and \( \text{Supp}(\lor AB) \neq \text{Supp}(\lor ABE) \). Indeed, \( D \) and \( E \) appear in the eighth transaction where A and B do not appear. Thus, only C can augment AB without affecting its disjunctive support. Consequently, \( h(AB) = ABC \).

3.2 Properties

In the following, we present the main theoretical properties of the (compound) operators we introduced. The associated proofs are given in the Appendix.

**Proposition 2.** The following properties hold for all \( I, I_1, I_2 \in \mathcal{P}(\mathcal{F}) \) and \( T, T_1, T_2 \in \mathcal{P}(\mathcal{T}) \):

Please cite this article in press as: T. Hamrouni et al., Sweeping the disjunctive search space towards mining new exact concise representations of frequent itemsets, Data Knowl. Eng. (2009), doi:10.1016/j.datak.2009.05.001
Proposition 5. Following proposition shows how to select the disjunctive closure of an arbitrary itemset.

Thus, AB is not a disjunctive closed itemset since C appears in the set of transactions where at least an item of AB appears, and nowhere else. On the other hand, the disjunctive support augments proportionally to itemset sizes, i.e., $\text{Supp}(I_1) < \text{Supp}(I_2)$ if $I_1 \subseteq I_2$. Thus, it is sufficient to only compare the disjunctive support of I with those of its immediate supersets, instead of all, to check whether it is a disjunctive closed itemset or not. Let us give some examples of the closure operator $h$ that will be at the roots of the concise representation we will introduce.

Example 8. Given the dataset depicted by Table 1, the itemset CE is a disjunctive closed itemset, since it is equal to the largest set of items only contained in the set of transactions where C or E appears, i.e., $\{1, 2, 4, 6, 7, 8\}$. Hence, $h(CE) = CE$.

Using disjunctive supports, we have $\text{Supp}(\mathcal{CE}) = 6 < \min\{\text{Supp}(\mathcal{ACE}), \text{Supp}(\mathcal{BCE}), \text{Supp}(\mathcal{CDE})\} = 7$. On the other hand, AB is not a disjunctive closed itemset since C only appears in the set of transactions where at least an item of AB appears. Actually, $h(AB) = ABC$.

In the remainder, the set of all disjunctive closed itemsets that can be drawn from a dataset $\mathcal{D}$ will be denoted $\mathcal{DCI}$. The following proposition shows how to select the disjunctive closure of an arbitrary itemset $I$ among those belonging to $\mathcal{DCI}$.

Proposition 5. Let $I \subseteq \mathcal{I}$. The itemset $h(I)$ is the smallest disjunctive closure containing $I$: $h(I) = \min\{I_1 \in \mathcal{DCI} | I_1 \subseteq I\}$.

Proof. The proof straightforwardly derives from the definition of a disjunctive closed itemset.

Proposition 6 establishes the link between the disjunctive support of an itemset and that of its closure.

Proposition 6. Let $I \subseteq \mathcal{I}$. $\text{Supp}(\bigvee I) = \text{Supp}(\bigvee h(I))$.

Proof. According to Property (4') (cf., Proposition 2), we have $g(I) = g \circ f \circ g[I]$. Hence, $g(I) = g[h(I)]$. We then have: $\text{Supp}(I) = |g(h(I))|$. It follows that $\text{Supp}(\bigvee I) = \text{Supp}(\bigvee h(I))$. 

The following two propositions straightforwardly derive from Proposition 2.

Proposition 3. Operator $h$ is a closure operator.

Proof. According to Proposition 2, $h$ is extensive (cf. Property (2)), isotonous (cf. Property (3)) and idempotent (cf. Property (5)). Thus, operator $h$ is a closure operator.

Proposition 4. Operator $h'$ is a kernel operator.

Proof. According to Proposition 2, $h'$ is contractive (cf. Property (2')), isotonous (cf. Property (3')) and idempotent (cf. Property (5')). Thus, operator $h'$ is a kernel operator.
Proposition 7 shows that it is possible to deduce the disjunctive closure of an itemset thanks to one of its subsets.

Proposition 7. Let \( I, I_1 \subseteq \mathcal{I} \) be two itemsets. We then have:

\[
(I \subseteq I \subseteq h(I_1)) \Rightarrow (h(I) = h(I_1)).
\]

**Proof.** We have \( I_1 \subseteq I \subseteq h(I_1) \). Since \( h \) is isotone as being a closure operator, we obtain \( h(I_1) \subseteq h(I) \subseteq h(h(I_1)) \). Thanks to the idempotency property, we get \( h(I_1) \subseteq h(h(I)) \subseteq h(I_1) \). Thus, we can conclude that \( h(I) = h(I_1) \). \( \square \)

Thanks to Proposition 8, we establish the link between disjunctive closed itemsets and essential itemsets.

Proposition 8. Let \( \mathcal{E} \mathcal{I} \) be the set of all essential itemsets that can be extracted from a dataset \( \mathcal{D} \).

\[
\forall (I \subseteq \mathcal{I}), \exists (I_1 \in \mathcal{D} \mathcal{E} \mathcal{I} \text{ and } I_2 \in \mathcal{E} \mathcal{I}) \text{ such that } h(I_1) = h(I) = I_1 \text{ and } I_2 \subseteq I.
\]

**Proof.** Let \( X \in \mathcal{E} \mathcal{I} \) be a maximal subset of \( I \) such that \( \text{Supp}(\cup X) = \text{Supp}(\cup I) \). Hence, \( g(X) = g(I) \). By applying \( f \), we have:

\[
f \circ g(X) = f \circ g(I).
\]

Hence, \( h(X) = h(I) \). Since \( I \subseteq h(I) \), then \( I \subseteq h(X) \). We can then conclude that there is a disjunctive closed itemset \( I_1 = h(X) \) associated to an essential itemset, namely \( X \), that contains \( I \). It is hence sufficient to take \( I_2 = X \). \( \square \)

It is important to mention that Propositions 6 and 7 offer a new characterization of essential itemsets. Indeed, their original characterization was based on their associated supports (cf. Definition 4). The new characterization, based on disjunctive closed itemsets, is as follows:

Proposition 9. Let \( I \subseteq \mathcal{I} \). \( I \) is an essential itemset if \( \forall I_1 \subset I, I \not\subseteq h(I_1) \).

**Proof.** Suppose that \( \exists I_1 \subset I \text{ st. } I \subseteq h(I_1) \). According to Proposition 7, we have \( h(I) = h(I_1) \). Thanks to Proposition 6, we have \( \text{Supp}(\cup h(I_1)) = \text{Supp}(\cup h(I)) = \text{Supp}(\cup I) \). Since \( \text{Supp}(\cup I_1) = \text{Supp}(\cup I) \), then \( I \) is not an essential itemset. Thus, if \( I \) is an essential itemset, then \( \forall I_1 \subset I, I \not\subseteq h(I_1) \). \( \square \)

The following proposition ensures that it is possible to derive the disjunctive support of each subset of an arbitrary itemset starting from the set \( \mathcal{D} \mathcal{E} \mathcal{I} \) of disjunctive closed itemsets.

Proposition 10. Let \( I \subseteq \mathcal{I} \). \( I \subseteq \mathcal{I} \), the disjunctive support of \( I_1 \) can be exactly derived from the set \( \mathcal{D} \mathcal{E} \mathcal{I} \) of disjunctive closed itemsets.

**Proof.** The set \( \mathcal{D} \mathcal{E} \mathcal{I} \) contains all the disjunctive closed itemsets that can be drawn from a dataset \( \mathcal{D} \). Hence, \( \forall I \subseteq \mathcal{I} \), \( h(I_1) \in \mathcal{D} \mathcal{E} \mathcal{I} \). We can thus retrieve the exact disjunctive support of \( I_1 \) thanks to Proposition 6. \( \square \)

5. Disjunctive closure-based concise representations

### 5.1. New concise representation for all itemsets

Let us begin by introducing a concise representation of the whole set of itemsets based on disjunctive closed itemsets. This is stated in Theorem 1.

**Theorem 1.** The set \( \mathcal{D} \mathcal{E} \mathcal{I} \) of disjunctive closed itemsets, associated to their respective disjunctive supports, is an exact concise representation of the whole set of itemsets.

**Proof.** Let \( I \subseteq \mathcal{I} \). It was proven through Proposition 10 that the disjunctive support of \( I \) and those of its subsets can be exactly derived from \( \mathcal{D} \mathcal{E} \mathcal{I} \). Then, by applying an inclusion–exclusion identity using the obtained disjunctive supports (cf. Lemma 1), we are able to get the exact conjunctive support of \( I \). \( \square \)

The set \( \mathcal{D} \mathcal{E} \mathcal{I} \) is thus not only a concise representation of frequent itemsets but also that of the whole set of itemsets that can be drawn from a dataset (i.e., even the associated supports of infrequent itemsets can be derived using \( \mathcal{D} \mathcal{E} \mathcal{I} \)).

### 5.2. Effect of setting the conjunctive frequency constraint

In practice, end-users are mainly interested in frequent itemsets and not in all itemsets. The selection of frequent itemsets can be done as a post-treatment by comparing the obtained supports with \( \text{minsupp} \). Nevertheless, it is more advantageous to restrict the representation to only the required elements while preserving the exact regeneration of frequent itemsets. Among these elements, disjunctive closed itemsets having at least a frequent essential itemset as generator should obviously be maintained. Indeed, they cover at least a frequent itemset, namely the associated frequent essential itemset. These closed sets, along with their associated disjunctive supports, will hence constitute the key information allowing to derive the exact disjunctive and, hence, conjunctive supports of frequent itemsets. The subset of \( \mathcal{D} \mathcal{E} \mathcal{I} \) containing these closures will be denoted \( \mathcal{D} \mathcal{E} \mathcal{F} \) (essential disjunctive closed itemsets). This set is then as follows:

Please cite this article in press as: T. Hamrouni et al., Sweeping the disjunctive search space towards mining new exact concise representations of frequent itemsets, Data Knowl. Eng. (2009), doi:10.1016/j.datak.2009.05.001
Definition 9. The set $\mathcal{EDCI}$ is equal to: $\mathcal{EDCI} = \{ h(I) \in \mathcal{F} | l \in \mathcal{FEI} \}$.

Example 9. For $\text{minsupp} = 1$, ABCDE $\in \mathcal{EDCI}$, since it has AD for frequent essential itemset.

The next lemma compares the size of $\mathcal{EDCI}$ with that of $\mathcal{FEI}$.

Lemma 2. The cardinality of $\mathcal{EDCI}$ is at most equal to that of $\mathcal{FEI}$.

Proof. To each frequent essential itemset is associated a unique element in $\mathcal{EDCI}$. Hence, the size of the set $\mathcal{EDCI}$ will be lower than or equal to that of $\mathcal{FEI}$. □

Thanks to Lemma 3, we can correctly derive the disjunctive supports of frequent itemsets from the elements of $\mathcal{EDCI}$. Once disjunctive supports derived, Lemma 1 will then be used when desired to deduce their conjunctive and negative supports.

Lemma 3. Let $\mathcal{F}$ be the set of frequent itemsets, $I \subseteq \mathcal{F}$ and $I_{min} = \min \{ I \in \mathcal{EDCI} | l \subseteq I \}$ if it exists. We then have:

$$\forall I \in \mathcal{F}, (\exists I_{min}) \land (\text{Supp}(\lor I) = \text{Supp}(\lor I_{min})).$$

Proof. The proof straightforwardly derives from that of Proposition 8 and the fact that the disjunctive closure of a frequent itemset $I$ is the smallest one covering it among those of $\mathcal{EDCI}$. □

As mentioned above, a concise representation of frequent itemsets based on disjunctive closed itemsets must contain the elements of $\mathcal{EDCI}$. Nevertheless, is this set sufficient to offer an exact concise representation?

During the regeneration process of frequent itemsets, the minimal infrequent itemsets are also checked since having all their subsets frequent. Let $I$ be such an itemset. If $I$ is not covered by any closure of $\mathcal{EDCI}$, then it is infrequent according to Lemma 3. However, the itemset $I$ can be covered by an element belonging to $\mathcal{EDCI}$, while having its proper closure not in $\mathcal{EDCI}$. Indeed, recall that the set $\mathcal{EDCI}$ results from combining two constraints of different types, namely a monotone one through the disjunctive support and an anti-monotone constraint using $\text{minsupp}$. Some key disjunctive closed itemsets for a correct regeneration process may thus be pruned since having all their essential itemsets infrequent. This leads to affecting to $I$ a wrong disjunctive support which, in some cases, will incorrectly make $I$ frequent instead of infrequent. Let us take a concrete example.

Example 10. Let us consider the dataset shown in Table 1 for $\text{minsupp} = 1$. Let $I = \text{CDE}$. The immediate subsets of CDE, namely CD, CE and DE, are frequent essential itemsets. They are respectively equal to their associated disjunctive closures.

During the regeneration process, the itemset CDE will be checked since having all its proper subsets frequent. Suppose that we do not maintain its proper closure in the representation and we will search for the smallest closure covering it in the representation. This latter will be equal to ABCDE, since one of its essential itemset, namely AD, is frequent. The closure of CDE will then be wrongly considered as being equal to ABCDE. It follows that its disjunctive support will be considered equal to 8 instead of 7. The conjunctive support of CDE will be derived as being equal to 1 (instead of 0). Then, CDE will be incorrectly considered as frequent, while it is actually infrequent.

The previous example clearly shows that $\mathcal{EDCI}$ cannot constitute by itself an exact concise representation of frequent itemsets. We thus need to retain some closures, in addition to $\mathcal{EDCI}$, that ensure correctly flagging the frequency status of itemsets whenever a wrong computation can arise. The following subsection explores this issue.

5.3. New concise representation of frequent itemsets

We propose here a new concise representation of frequent itemsets based on disjunctive closed itemsets. This representation is obtained by only adding some disjunctive closures to $\mathcal{EDCI}$. In this respect, the added itemsets constitute the set $\mathcal{ADCI}$ (stands for Added Disjunctive Closed Itemsets). This set is defined as follows:

Definition 10. Let $\mathcal{E}$ be the set of all essential itemsets that can be extracted from a dataset $\mathcal{D}$. The set $\mathcal{ADCI}$ is defined as follows: $\mathcal{ADCI} = \{ h(I) \in \mathcal{F} | l \in \mathcal{E} \land (\mathcal{F} \cup \mathcal{E}) \land ((-1)^{|I|} = -1) \land (\lor I_1 \subseteq \mathcal{F}, h(I_1) = h(I) \Rightarrow \text{Supp}(I_1) < \text{minsupp}) \}$.

Roughly speaking, $\mathcal{ADCI}$ is the set of the disjunctive closed itemsets, generated by infrequent essential itemsets of odd sizes and belonging to the negative border of frequent essential itemsets. These closures must have all their corresponding essential itemsets as infrequent.

Example 11. For $\text{minsupp} = 1$, CDE $\in \mathcal{ADCI}$. Indeed, its unique essential itemset is itself. Moreover, CDE is an odd-sized infrequent itemset, having all its proper subsets frequent. It hence belongs to the negative border of frequent essential itemsets. The next proposition states that both sets $\mathcal{EDCI}$ and $\mathcal{ADCI}$ are disjoint.

Please cite this article in press as: T. Hamrouni et al., Sweeping the disjunctive search space towards mining new exact concise representations of frequent itemsets, Data Knowl. Eng. (2009), doi:10.1016/j.datak.2009.05.001
Proposition 11. We have: \( \text{ADCI} \cap \text{DCI} = \emptyset \).

Proof. The proof is based on the fact that each element of \( \text{ADCI} \) has at least a frequent essential itemset as a seed, while all the essential itemsets of an element belonging to \( \text{DCI} \) are infrequent. \( \square \)

In the remainder, DCIs_rep stands for the concise representation \( \text{EDCI} \cup \text{ADCI} \). The exactness of the representation based on DCIs_rep is provided through Theorem 2.

Theorem 2. The set DCIs_rep of disjunctive closed itemsets, associated to their respective disjunctive supports, is an exact concise representation of the set \( \mathcal{F} \) of frequent itemsets.

Proof. Let \( I \subseteq \mathcal{F} \). If \( \exists I \subset I \) s.t. \( I \) is infrequent, then \( I \) is also infrequent. Otherwise (i.e., \( \forall 1 \subset I, I_1 \in \mathcal{F} \)), we need to show that the frequency status of \( I \) is correctly retrieved starting from DCIs_rep. In addition, its conjunctive support must be exactly computed if it is frequent. Two cases have to be distinguished:

1. If \( I \) is frequent, then its conjunctive support will be correctly derived thanks to Lemma 3 (cf. p. 11). Indeed, \( I \) is either a frequent essential itemset and its closure is in \( \text{ADCI} \), or encompassed between a frequent essential itemset and its closure, obviously belonging to \( \text{EDCI} \). Once its disjunctive support at hand, the computation of the conjunctive support becomes then straightforward thanks to an inclusion–exclusion identity.

2. If \( I \) is infrequent, then two cases arise:
   (a) If \( I \) is not an essential itemset, then it is contained in the disjunctive closure of one of its subsets. By hypothesis, this latter is frequent and hence its closure belongs to \( \text{ADCI} \). Also in this case, the disjunctive support of \( I \) will be correctly derived and hence its conjunctive support. By comparing the conjunctive support of \( I \) with \( \minsupp \), we get the information that \( I \) is infrequent.
   (b) If \( I \) is an essential itemset, then necessarily \( I \in \text{Ad}^{-1}(\mathcal{F}) \cap \mathcal{F} \). Let \( h_i \) be the smallest disjunctive closed itemset in \( \text{ADCI} \) containing \( I \). If \( h_i \) does not exist then \( I \) is immediately guessed to be infrequent (thanks to Lemma 3). Otherwise, from Formula (1) of Lemma 1, we have:

\[
\text{Supp}(I) = \sum_{\emptyset \subset C \subseteq I} (-1)^{|C|-1}\text{Supp}(\{I\}) = \sum_{\emptyset \subset C \subseteq I} (-1)^{|C|-1}\text{Supp}(\{I\}) + \sum_{\emptyset \subset C \subseteq I} (-1)^{|C|-1}\text{Supp}(\{I\}).
\]

Hence, according to the size of \( I \) we have: i. If \( |I| \) is even, then \( \text{Supp}(I) = \text{Supp}(\emptyset) + \sum_{\emptyset \subset C \subseteq I} (-1)^{|C|-1}\text{Supp}(\{I\}) < \minsupp \) (since \( I \) is infrequent). Since \( I \subseteq h_i \), we have \( \text{Supp}(\emptyset) = \text{Supp}(\emptyset) \). Hence, \( -\text{Supp}(h_i) + \sum_{\emptyset \subset C \subseteq I} (-1)^{|C|-1}\text{Supp}(\{I\}) = -\text{Supp}(h_i) + \sum_{\emptyset \subset C \subseteq I} (-1)^{|C|-1}\text{Supp}(\{I\}) < \minsupp \). This inequality points out that even if \( h_i \) is not necessarily the disjunctive closure of \( I \), we can detect that \( I \) is infrequent.

ii. If \( |I| \) is odd, then by applying the same process as for the previous case, we are not able to detect in all the cases the frequency status of \( I \). Indeed, in this case, \( (-1)^{|C|-1}\text{Supp}(\{I\}) \) is a positive quantity and not a negative one as in the case where \( |I| \) is even. Hence, if \( h_i \) is not the correct closure of \( I \), then \( h(i) \) has all its essential itemsets infrequent. It then belongs to \( \text{ADCI} \) (cf. Definition 10) and its addition to the representation is necessary to ensure the correct detection of the status of \( I \).

Thus, the set DCIs_rep is an exact concise representation of \( \mathcal{F} \). \( \square \)

The proof of Theorem 2 can easily be transformed to a naive algorithm for deriving frequent itemsets and their associated supports starting from our representation. Indeed, this can straightforwardly be done in a levelwise manner that regenerates 1-frequent itemsets, 2-frequent itemsets, and so forth.

It is important to mention that in the definition of the set \( \text{ADCI} \), we pruned the disjunctive closures of infrequent non-essential itemsets belonging to \( \text{Ad}^{-1}(\mathcal{F}) \) since they are already included in \( \text{EDCI} \) (cf. Proposition 7). On the other hand, we can further reduce the cardinality of \( \text{ADCI} \), and consequently that of DCIs_rep, by withdrawing its elements which are not covered by any disjunctive closed itemset of \( \text{EDCI} \). Thus, hereafter, \( \text{ADCI} \) will denote the set presented in Definition 10, from which we prune its elements not covered by at least a closure of \( \text{EDCI} \). Hence, \( \text{ADCI} = \text{ADCI} \setminus \{I \in \text{ADCI} | \exists I_1 \in \text{EDCI} s.t. I \subset I_1\} \).

The next theorem states the correctness of the regeneration process of frequent itemsets after this pruning.

Theorem 3. After pruning non-covered itemsets from \( \text{ADCI} \), the set DCIs_rep is still an exact concise representation of the set \( \mathcal{F} \) of frequent itemsets.

Proof. The proof is based on that of Theorem 2 and on Lemma 3. Indeed, if an itemset \( I \) is not covered by any element of \( \text{ADCI} \), then we can directly assert that \( I \) is infrequent (cf. Lemma 3). Therefore, thanks to the extensivity property of any closure operator, \( h(I) \) cannot be subsumed by any element of \( \text{EDCI} \). Thus, it can be pruned from \( \text{ADCI} \) while ensuring the correctness of the regeneration mechanism (cf. proof of Theorem 2). Thus, after pruning non-covered itemsets from \( \text{ADCI} \), the set DCIs_rep is still an exact concise representation of \( \mathcal{F} \). \( \square \)
The different representations for the dataset in Table 1 and for \( \text{minsupp} = 1 \).

<table>
<thead>
<tr>
<th>Size</th>
<th>DCIs_rep</th>
<th>FCI_rep</th>
<th>FDI_rep</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(A,6) (B,6) (C,3) (D,4) (E,4)</td>
<td>(A,6) (B,6) (C,3) (D,4) (E,4)</td>
<td>( \mathcal{F} )</td>
</tr>
<tr>
<td>2</td>
<td>(A,6) (B,6) (C,3) (D,4) (E,4)</td>
<td>(A,6) (B,6) (C,3) (D,4) (E,4)</td>
<td>( \mathcal{F} )</td>
</tr>
<tr>
<td>3</td>
<td>(A,6) (B,6) (C,3) (D,4) (E,4)</td>
<td>(A,6) (B,6) (C,3) (D,4) (E,4)</td>
<td>( \mathcal{F} )</td>
</tr>
<tr>
<td>4</td>
<td>(A,6) (B,6) (C,3) (D,4) (E,4)</td>
<td>(A,6) (B,6) (C,3) (D,4) (E,4)</td>
<td>( \mathcal{F} )</td>
</tr>
<tr>
<td>5</td>
<td>(A,6) (B,6) (C,3) (D,4) (E,4)</td>
<td>(A,6) (B,6) (C,3) (D,4) (E,4)</td>
<td>( \mathcal{F} )</td>
</tr>
</tbody>
</table>

Example 12. This example proposes a comparison of our representation DCIs_rep vs. the main representations of the literature in terms of size. Thus, Table 2 presents the different concise representations of frequent itemsets associated to the dataset shown in Table 1 and \( \text{minsupp} = 1 \). To this end, the abbreviation "FCIs_rep" (resp. "NDIs_rep", "CNDIs_rep", and "FEIs_rep") is used to stand for the frequent closed [40] (resp. non-derivable [12], closed non-derivable [38] and essential [15]) itemset-based concise representation. For each representation, its elements are presented using couples representing an itemset belonging to the representation and its associated conjunctive or disjunctive support according to its membership in \( \mathcal{F} \) or \( \mathcal{D} \).

For this dataset, the cardinalities of the different representations are respectively as follows: \(|\text{DCIs}_\text{rep}| = 13, \ |\text{FCIs}_\text{rep}| = 21, \ |\text{FEIs}_\text{rep}| = 17, \ |\text{NDIs}_\text{rep}| = 18, \text{and} \ |\text{CNDIs}_\text{rep}| = 16 \). Note also that, in this case, the size of \( \mathcal{F} \) is equal to 23.

Recall that to belong to \( \text{FCIs}_\text{rep} \), an itemset must be frequent and of conjunctive support strictly higher than those of its strict supersets. While an itemset belonging to \( \text{NDIs}_\text{rep} \) must be frequent, and having a conjunctive support not exactly derivable using the deduction rules based on the conjunctive supports of all its subsets [12]. The main advantage of \( \text{NDIs}_\text{rep} \) is then brought by the large neighborhood explorations to retain or not an itemset within the representation. The concise representation \( \text{CNDIs}_\text{rep} \) simply gathers the conjunctive closures of frequent non-derivable itemsets. Consequently, its cardinality is always smaller than or equal to those of \( \text{NDIs}_\text{rep} \) and \( \text{FCIs}_\text{rep} \). This is also confirmed by the obtained results in this comparison although the difference between \( \text{CNDIs}_\text{rep} \) and \( \text{FCIs}_\text{rep} \) only consists in the underlined element in \( \text{FCIs}_\text{rep} \), ABDE. This latter closure does not have any non-derivable itemset as generator, and is hence not retained in \( \text{CNDIs}_\text{rep} \).

It is important to mention that although non-derivable itemsets exploit much larger neighborhoods (for each candidate, all its subsets are used), the size of the associated representation is greater than that based on disjunctive closed itemsets. Interestingly, taking their closure to obtain \( \text{CNDIs}_\text{rep} \) only allows eliminating a unique element, namely ABDE. This clearly shows that exploring the disjunctive search space offers new opportunities for a further reduction of the retained itemsets. Moreover, computing (closed) non-derivable itemsets is awfully costly since it requires the evaluation of \( 2^n \) deduction rules for each itemset of size \( n \) (in addition to taking closure if \( \text{CNDIs}_\text{rep} \) is the targeted representation) [12,38].

It is interesting to note that for the concise representation \( \text{FEIs}_\text{rep} \), the itemsets ACE and BCD simultaneously belong to \( \mathcal{F} \) and \( \mathcal{D}(\mathcal{F}) \) since being frequent essential itemsets and maximal frequent itemsets. In this respect, note that \( \mathcal{D}(\mathcal{F}) \subseteq \text{FCIs}_\text{rep} \) since a maximal frequent itemset is obviously closed.

On its side, for the representation \( \text{DCIs}_\text{rep} \), a unique element is underlined to indicate that it belongs to \( \mathcal{F} \). The other closures form \( \mathcal{E} \). Note that for 17 frequent essential itemsets, we only have 12 disjunctive closed itemsets in \( \mathcal{E} \). In addition, four maximal frequent itemsets are required for ensuring an exact regeneration process of frequent itemsets starting from \( \text{FEIs}_\text{rep} \).

From the point of view of equivalence classes, it is easily observable that disjunctive equivalence class gathers itemsets do not having necessarily the same conjunctive support. For example, AD and ABCDE belong to the same disjunctive equivalence class and, hence share the same conjunctive support equal to 8. However, they have different conjunctive support equal to 2 for the former and 0 for the latter. It is the reverse for conjunctive equivalence classes. On the other hand, frequent non-derivable itemsets do not represent the different conjunctive equivalence classes. Indeed, the equivalence class whose the frequent closed itemset is ABDE does not have any of its members being non-derivable. For this reason, ABDE \( \notin \text{CNDIs}_\text{rep} \).

If we look for the intersection between the different representations, we find that A, B, C, D, E, AE, BD and DE are simultaneously disjunctive/conjunctive closed, non-derivable and essential itemsets.
5.4. Features of the proposed representation

In addition to the exact retrieval of frequent itemsets as well as their different kinds of supports, DCIs_rep presents the following main properties:

1. **Homogeneity**: The set dcis_rep overcomes the heterogeneity problem since it only involves disjunctive closed itemsets (vs., for example, $\mathcal{F}$ and $\mathcal{F}^d(\mathcal{F})$). Its elements hence have the same structural properties. Indeed, they are the top elements of their associated equivalence classes within the disjunctive search space. This ensures the disjunctive support.

2. **Redundancy free**: Redundancy is due to the fact that a set of disjunctive itemsets can characterize the same set of transactions. This is avoided in our case since such a set is simply represented by a unique disjunctive closed itemset, thanks to the proposed disjunctive closure operator.

3. **Mining algorithm**: The disjunctive closed itemsets composing the dcis_rep representation have, for associated seeds, the set $\mathcal{F}$ of frequent essential itemsets and a subset of the infrequent part of the associated negative border. Interestingly, these elements form a downward closed set. Thus, a levelwise (or a depth-first right-to-left) traversal of the search space is indicated for localizing the aforementioned elements without overhead w.r.t. those of the negative border [35]. Indeed, the negative border consists of exactly those itemsets which, on the basis of other information, could be frequent essential, and on which the constraint "to be frequent essential" should therefore be checked.

4. **Small size**: $\mathcal{F}$ is the smallest set that concisely represents the equivalence classes containing at least a frequent itemset, since only a unique element is maintained per class. In addition, the size of $\mathcal{F}$ is expected to be very small compared to $\mathcal{F}^d(\mathcal{F})$, since its elements must fulfill many easy-to-check constraints. This will be confirmed by experiments where DCIs_rep is shown to provide very interesting compactness rates.

5. **Low regeneration cost**: It is worth mentioning that our concise representation allows retrieving the conjunctive support faster than from frequent non-derivable itemsets. Indeed, for an itemset $I$ of size $n$, the retrieval process of Supp($I$) from this representation requires the costly evaluation of $2^n$ deduction rules based on Bonferroni-inequalities [37]. The computation cost for inferring supports is then awfully high which makes this representation not very practical [30,37]. Note also that taking closures of frequent non-derivable itemsets to obtain the closed non-derivable representation complicates both the extraction process of this latter, as well as the regeneration process of frequent itemsets. On its side, the frequent closed itemset-based representation [40] allows retrieving the conjunctive support of $I$ by searching for the smallest closure containing it. However, it does not allow the straightforward derivation of its disjunctive and negative supports. While the retrieval of Supp($I$) from our concise representation only needs to evaluate a unique inclusion–exclusion identity. Moreover, given at hand Supp($I_1$) such that $I_1$ is an immediate subset of $I$ and $I \setminus I_1 = \emptyset$, we can straightforwardly deduce the support of $I$. Indeed, it derives from Formula (1) in Lemma 1 that:

$$\text{Supp}(I) = \text{Supp}(I_1) + \sum_{I_j \subseteq I \setminus I_1} (-1)^{|I_j|-1} \text{Supp}(\bigvee I_j).$$

6. Experimental results

In this section, our objective is to show, through extensive experiments, that our concise representation provides interesting compactness rates compared respectively to the representations based on frequent closed itemsets, frequent (closed) non-derivable itemsets and frequent essential itemsets. For this purpose, we implemented an algorithm for mining the disjunctive closed itemset-based representation. The source codes allowing the extraction of the remaining representations are kindly provided by their respective authors. For this purpose, we implemented an algorithm for mining the disjunctive closed itemset-based representation. The source codes allowing the extraction of the remaining representations are kindly provided by their respective authors.

The experiments were carried out on benchmark datasets whose characteristics are summarized in Table 3. The first five datasets are commonly considered to be dense (i.e., containing many long frequent itemsets at various levels of minsupp values [6]), while the last five are considered to be sparse (i.e., containing a large number of items but only a few of them frequently occur in the dataset). The connect dataset contains all legal 8-ply positions in the game of connect-4, while chess is derived from the steps of Chess games. The mushrooms dataset includes descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms. The pumsb dataset contains census data from PUMS (Public Use Microdata Samples), while pumsb* is obtained after deleting all frequent items for a minsupp value set to 80% in the original pumsb dataset. The accident dataset reports the traffic accidents from the National Institute of Statistics (NIS) for the region of Flanders (Belgium). The kosarak dataset contains data corresponding to (anonymized) click-stream data of a Hungarian on-line news portal. The retail dataset contains information

---

Please cite this article in press as: T. Hamrouni et al., Sweeping the disjunctive search space towards mining new exact concise representations of frequent itemsets, Data Knowl. Eng. (2009), doi:10.1016/j.datake.2009.05.001
about the market basket of clients in a Belgian supermarket. The T10I4D100K and T40I10D100K datasets are synthetic, generated using the generator from the IBM Almaden Quest research group.

All experiments were carried out on a PC equipped with a 1.73 GHz Intel processor and 2 GB of main memory, running the GNU/Linux distribution Fedora Core 7 (with 2 GB of swap memory).

Tables 4 and 6 compare the size of DCIs\_rep (resp. FEIs\_rep) vs. that of FEI (resp. \( A_d^* (F) \)) on dense and sparse datasets, respectively. Both tables also compare the size of the resulting representations, i.e., DCIs\_rep vs. FEIs\_rep. Note that the symbol ’/’ indicates that a ratio cannot be computed, since the size of ADCI is equal to 0. While Figs. 1 and 3 graphically sketch the obtained results for dense and sparse datasets, respectively. On the other hand, Tables 5 and 7 compare the size of our concise representation DCIs\_rep to those of the literature as well as to the size of the set of frequent itemsets, respectively on dense and sparse datasets. In both tables, we use the symbol ’-’ to designate a case where an execution error occurred. For example, to show the cardinality of CNDIs\_rep, the authors of [38] have chosen a specific interval of \( \text{minsupp} \) values for some datasets also used in our tests. Nevertheless, beyond these intervals, we noticed that their program comes to an end with an execution error. Obtained results are also graphically sketched by Fig. 2 and 4 for dense and sparse datasets, respectively. Note that we only selected figures with different layouts.

In the literature dedicated to concise representations of frequent itemsets (e.g., [9,13]), it was shown that dense datasets present the most interesting cases. Indeed, within such datasets, the compactness ratio between the cardinality of the set of frequent itemsets and those of concise representations is high. On the contrary, equivalence classes extracted from sparse datasets are often reduced to the associated generators and cannot be further compacted. The number of extracted frequent itemsets is hence small even for low \( \text{minsupp} \) values. This makes the size reduction rates brought by concise representations meaningless in such datasets. The next paragraphs give a thorough analysis of the obtained results.

---

**Table 3**

Dataset characteristics.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Nature</th>
<th># items</th>
<th># Transactions</th>
<th>Avg. size of transactions</th>
</tr>
</thead>
<tbody>
<tr>
<td>CONNECT</td>
<td>Dense</td>
<td>129</td>
<td>67,557</td>
<td>43.00</td>
</tr>
<tr>
<td>MUSHROOM</td>
<td>Dense</td>
<td>119</td>
<td>8,124</td>
<td>23.00</td>
</tr>
<tr>
<td>CHESS</td>
<td>Dense</td>
<td>75</td>
<td>3,196</td>
<td>37.00</td>
</tr>
<tr>
<td>PUMSB</td>
<td>Dense</td>
<td>2,113</td>
<td>49,046</td>
<td>74.00</td>
</tr>
<tr>
<td>PUMSB*</td>
<td>Sparse</td>
<td>2,088</td>
<td>49,046</td>
<td>50.48</td>
</tr>
<tr>
<td>ACCIDENTS</td>
<td>Sparse</td>
<td>468</td>
<td>340,183</td>
<td>33.81</td>
</tr>
<tr>
<td>KOSARAK</td>
<td>Sparse</td>
<td>41,270</td>
<td>990,002</td>
<td>8.10</td>
</tr>
<tr>
<td>RETAIL</td>
<td>Sparse</td>
<td>16,470</td>
<td>88,162</td>
<td>10.31</td>
</tr>
<tr>
<td>T10I4D100K</td>
<td>Sparse</td>
<td>1,000</td>
<td>100,000</td>
<td>10.10</td>
</tr>
<tr>
<td>T40I10D100K</td>
<td>Sparse</td>
<td>1,000</td>
<td>100,000</td>
<td>39.61</td>
</tr>
</tbody>
</table>

**Table 4**

Size of DCIs\_rep vs. FEI, and DCIs\_rep vs. \( A_d^* (F) \) for dense datasets.

<table>
<thead>
<tr>
<th>minsupp</th>
<th>DCIs_rep</th>
<th>FEIs_rep</th>
<th>FEI</th>
<th>( A_d^* (F) )</th>
<th>Ratios</th>
</tr>
</thead>
<tbody>
<tr>
<td>CONNECT</td>
<td>90</td>
<td>22</td>
<td>0</td>
<td>176</td>
<td>8.00</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>819</td>
<td>243</td>
<td>1,948</td>
<td>2.38</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>70,798</td>
<td>11,940</td>
<td>75,345</td>
<td>1.06</td>
</tr>
<tr>
<td>MUSHROOM</td>
<td>40</td>
<td>79</td>
<td>12</td>
<td>110</td>
<td>1.39</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>17,814</td>
<td>2,740</td>
<td>22,965</td>
<td>1.29</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>186,273</td>
<td>10,782</td>
<td>230,474</td>
<td>1.24</td>
</tr>
<tr>
<td>CHESS</td>
<td>90</td>
<td>40</td>
<td>3</td>
<td>84</td>
<td>2.10</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>1,695</td>
<td>276</td>
<td>2,809</td>
<td>1.66</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>70,354</td>
<td>5,844</td>
<td>114,219</td>
<td>1.62</td>
</tr>
<tr>
<td>PUMSB</td>
<td>90</td>
<td>263</td>
<td>55</td>
<td>529</td>
<td>2.01</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>5,316</td>
<td>234</td>
<td>17,256</td>
<td>3.25</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>487,186</td>
<td>13,319</td>
<td>2,182,183</td>
<td>4.48</td>
</tr>
<tr>
<td>PUMSB*</td>
<td>70</td>
<td>17</td>
<td>0</td>
<td>17</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>1,445</td>
<td>138</td>
<td>1,946</td>
<td>1.35</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>513,639</td>
<td>14,329</td>
<td>1,397,665</td>
<td>2.72</td>
</tr>
</tbody>
</table>

Please cite this article in press as: T. Hamrouni et al., Sweeping the disjunctive search space towards mining new exact concise representations of frequent itemsets, Data Knowl. Eng. (2009), doi:10.1016/j.datak.2009.05.001
For the different minsupp values, the compactness rate on dense datasets offered by our concise representation DCIs rep, w.r.t. the size of \( F \), is considerably high (cf. Table 5). For example, it reaches almost 1,116,705 times for CONNECT with minsupp = 20%. This clearly shows the necessity to set up concise representations for such type of datasets. The size of DCIs rep is also lower than or equal to that of \( F \) for sparse datasets. However, the obtained results confirm that the compactness rates offered by the pioneer concise representations of the literature are often low on such datasets. Indeed, the size of DCIs rep is almost equal to that of \( F \) for different sparse datasets, such as KOSARAK and T10I4D100K. This makes the associated curves collapse (cf. Fig. 4 (left)). Interestingly, for the ACCIDENTS dataset, we note a reduction reaching 8.23 for minsupp = 20% (cf. Table 7).

- DCIs rep vs. FEI rep: For the different benchmark datasets, the size of DCIs rep is always smaller than that of FEI rep (cf. Tables 4 and 6). Considering Table 4, the cardinality of \( \text{EDCI} \) is always lower than that of \( \text{FEI} \). By comparing the respective cardinalities of \( \text{ADCI} \) and \( \text{BD}^+ (F) \), we note that the associated ratio reaches high values whenever the size of

| minsupp (%) | DCIs_rep | \( |F| \) | DCIs_rep | \( |F| \) | DCIs_rep | \( |F| \) | DCIs_rep | \( |F| \) | DCIs_rep | \( |F| \) |
|-------------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|
| CONNECT     |          |          |          |          |          |          |          |          |          |          |
| 90          | 22       | 1,233.09 | 158.50   | 9.05     | 8.05     |
| 40          | 1,062    | 320,070.86 | 225.40  | 1.95     | –        |
| 5           | 82,738   | -        | 343.07   | 1.10     | –        |
| MUSHROOM    |          |          |          |          |          |          |          |          |          |          |
| 40          | 91       | 6.22     | 1.54     | 1.60     | 1.29     |
| 5           | 20,554   | 182.71   | 0.62     | 0.56     | 0.32     |
| 11          | 197,055  | 460.54   | 0.26     | 0.25     | –        |
| CHESS       |          |          |          |          |          |          |          |          |          |          |
| 90          | 43       | 14.49    | 11.60    | 2.21     | 2.16     |
| 5           | 1,971    | 645.83   | 187.44   | 1.74     | 1.70     |
| 10          | 76,198   | 59,762.45 | 1,617.41 | 1.29     | –        |
| PUMSB       |          |          |          |          |          |          |          |          |          |          |
| 90          | 318      | 8.20     | 4.61     | 1.84     | 1.45     |
| 60          | 5,550    | 3,518.92 | 193.63   | 3.84     | 2.18     |
| 20          | 500,505  | –        | 14,891.96 | –        | –        |
| PUMSB*      |          |          |          |          |          |          |          |          |          |          |
| 70          | 17       | 1.76     | 1.06     | 1.24     | 1.06     |
| 40          | 1,593    | 17.17    | 1.64     | 1.18     | 1.00     |
| 10          | 527,968  | –        | 2.87     | 0.85     | –        |

Table 5
The compactness rates offered by the representation based on disjunctive closed itemsets for dense datasets.

Table 6
Size of \( \text{EDCI} \) vs. \( \text{FEI} \), and \( \text{EDCI} \) vs. \( \text{BD}^+ (F) \) for sparse datasets.

Please cite this article in press as: T. Hamrouni et al., Sweeping the disjunctive search space towards mining new exact concise representations of frequent itemsets, Data Knowl. Eng. (2009), doi:10.1016/j.datak.2009.05.001
Table 7

The compactness rates offered by the representation based on disjunctive closed itemsets for sparse datasets.

| minsupp (%) | DCIs_rep | | | EDCIs_rep | | | FIs | | | FEIs | | | NDCIs_rep | | | NNDIs_rep |
|-------------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|
| ACCIDENTS   |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |
| 50          | 2, 497   | 3.23     | 3.23     | 1.14     | 1.14     |          |          |          |          |          |          |          |          |          |          |
| 30          | 25, 588  | 5.84     | 5.84     | 1.12     | 1.12     |          |          |          |          |          |          |          |          |          |          |
| 20          | 108, 123 | 8.23     | 8.21     | 1.02     | 1.02     |          |          |          |          |          |          |          |          |          |          |
| KOSARAK     |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |
| 1.00        | 383      | 1.00     | 1.00     | 1.00     | 1.00     |          |          |          |          |          |          |          |          |          |          |
| 0.40        | 2, 522   | 1.00     | 1.00     | 1.00     | 1.00     |          |          |          |          |          |          |          |          |          |          |
| 0.20        | 39, 464  | 1.00     | 0.91     | 0.44     | 0.43     |          |          |          |          |          |          |          |          |          |          |
| RETAIL      |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |
| 1.00        | 159      | 1.01     | 1.01     | 1.01     | 1.01     |          |          |          |          |          |          |          |          |          |          |
| 0.10        | 7, 588   | 1.00     | 1.00     | 1.00     | 1.00     |          |          |          |          |          |          |          |          |          |          |
| 0.05        | 19, 238  | 1.00     | 0.99     | 1.00     | 0.99     |          |          |          |          |          |          |          |          |          |          |
| T10I4D100K  |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |
| 1.00        | 385      | 1.00     | 1.00     | 1.00     | 1.00     |          |          |          |          |          |          |          |          |          |          |
| 0.20        | 12, 950  | 1.02     | 1.01     | 0.89     | 0.88     |          |          |          |          |          |          |          |          |          |          |
| 0.02        | 127, 528 | 1.02     | 0.85     | 0.86     | 0.81     |          |          |          |          |          |          |          |          |          |          |
| T40I10D100K |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |
| 3           | 793      | 1.00     | 1.00     | 1.00     | 1.00     |          |          |          |          |          |          |          |          |          |          |
| 2           | 2, 293   | 1.00     | 1.00     | 1.00     | 1.00     |          |          |          |          |          |          |          |          |          |          |
| 1           | 65, 236  | 1.00     | 1.00     | 0.65     | 0.65     |          |          |          |          |          |          |          |          |          |          |

A $\text{ADCI}$ is smaller than that of $\text{Bd}^+ (\mathcal{F})$. This occurs for the CONNECT, CHESS and PUMSB datasets which explains why our representation is largely smaller than FEI_rep on these datasets. For MUSHROOM, $\text{ADCI}$ is smaller than $\text{Bd}^+ (\mathcal{F})$ for high minsupp values while it is the opposite for low values, although the ratio values are too small. It is the opposite for PUMSB w.r.t. minsupp values while preserving the low ratio values. On the other hand, by comparing the respective size of the couple of sets constituting each representation on sparse datasets, we notice that the size of $\text{EDCI}$ is equal to that of $\text{FEI}$ for the KOSARAK, RETAIL and T40I10D100K datasets. While its size is slightly reduced for T10I4D100K for very low minsupp values, and ACCIDENTS for all minsupp values. Consequently, for the KOSARAK and T10I4D100K datasets, the curves representing the size of the sets $\text{EDCI}$ and $\text{FEI}$ collapse (cf. Fig. 3 (top)). In fact, for sparse datasets, the main advantage of our representation is that it avoids the use of elements from the conjunctive search space contrary to FEIs_rep, which heavily relies on $\text{Bd}^+ (\mathcal{F})$. This border clearly increases the size of FEIs_rep. For example, the size of $\text{Bd}^+ (\mathcal{F})$ reaches 5, 025.80 and 13, 933.50 times the size of $\text{ADCI}$ for a minsupp value equal to 0.02% and 0.03% respectively (cf. Table 6). It is worth noting...
Fig. 2. Size of DCIs_rep vs. the whole set of frequent itemsets (left), vs. FEIs_rep (middle), and vs. the remaining representations (right) for dense datasets.

Fig. 3. Size of EDCIs vs. FEIs (top), and ADCIs vs. Bd+(FIs) (bottom) for sparse datasets.

Please cite this article in press as: T. Hamrouni et al., Sweeping the disjunctive search space towards mining new exact concise representations of frequent itemsets, Data Knowl. Eng. (2009), doi:10.1016/j.datak.2009.05.001
that this latter set is almost empty for all datasets, except \textit{accidents}. Indeed, its size for the other four datasets does not exceed 10. In the figure associated to the \textit{Kosarak} dataset, only the curve representing the size of $\textit{ndi}(\mathcal{F})$ appears (cf. Fig. 3 (Bottom)) since the size of $\textit{ndi}(\mathcal{F})$ is always equal to 0.

- \textit{DCIs\_rep vs. FCIs\_rep, NDI\_rep} and \textit{CNDIs\_rep}: For the \textit{Chess, Connect} and \textit{PUMS} datasets, the cardinality of \textit{DCIs\_rep} is significantly reduced compared to those of the other representations. It is also the case for the \textit{PUMS} dataset \textit{w.r.t.} \textit{FCIs\_rep} and \textit{FEIs\_rep}. However, for the \textit{Mushroom} dataset, the size of \textit{DCIs\_rep} is quite greater than the size of \textit{FCIs\_rep} and \textit{NDIs\_rep} for minsup values lower than 10%. For these datasets, \textit{DCIs\_rep} is, in most cases, less sensitive to the variation of minsup values than the other concise representations (cf. Fig. 2). Our representation is also the smallest one for the \textit{Accidents} dataset. For the remaining sparse datasets, its size is almost equal to that of \textit{FCIs\_rep} while it is greater than those of \textit{NDIs\_rep} and, consequently, \textit{CNDIs\_rep}. In this respect, it is important to note that the size of \textit{CNDIs\_rep} is almost equal to that of \textit{NDIs\_rep}. Hence, in such datasets, computing the closures associated to non-derivable itemsets to obtain \textit{CNDIs\_rep} is often useless, since each itemset is equal to its closure. This leads us to the following remark: in comparison to the large neighborhood explorations used for retaining or not an itemset within the \textit{NDIs\_rep} representation, the \textit{DCIs\_rep} only relies on taking closures of essential itemsets. These latter itemsets are based on a simple comparison of their Galois closure \cite{MC92} – within the conjunctive search space. An essential itemset is then the mapping of the concept of \textit{minimal generator} \cite{C&H97} (aka \textit{key} \cite{W97} and \textit{0-free} \cite{GK99} set). While a disjunctive closed itemset is the mapping of the concept of \textit{conjunctive closed itemset} \cite{YQ03}. A critical structural and analytical survey on mining algorithms for frequent closed itemsets can be found in \cite{DN99}.

7. Related work and discussion

First of all, let us make an alignment between the disjunctive and the conjunctive search spaces. We will hence find that disjunctive equivalence classes correspond to conjunctive equivalence classes – gathering itemsets having the same Galois closure \cite{MC92} – within the conjunctive search space. An essential itemset is then the mapping of the concept of \textit{minimal generator} \cite{C&H97} (aka \textit{key} \cite{W97} and \textit{0-free} \cite{GK99} set). While a disjunctive closed itemset is the mapping of the concept of \textit{conjunctive closed itemset} \cite{YQ03}. A critical structural and analytical survey on mining algorithms for frequent closed itemsets can be found in \cite{DN99}.

Please cite this article in press as: T. Hamrouni et al., Sweeping the disjunctive search space towards mining new exact concise representations of frequent itemsets, Data Knowl. Eng. (2009), doi:10.1016/j.datak.2009.05.001
The concepts of essential and disjunctive closed itemsets are closely related to many important pattern classes as detailed in the following. They can be considered as particular cases of composite items [60] where the disjunction of (infrequent) items is used to compose new items, the composite items. The work in [47] can be considered as an extension of composite items, since it takes into account particular disjunctive normal forms (DNFs) where disjuncts may contain a conjunction of items – frequent closed itemsets – and not only a single item. On the other hand, essential itemsets of the disjunctive equivalence class whose disjunctive support is equal to \( |P| \) are the minimal transversals of the hypergraph [19] represented through the mined dataset. In this latter, transactions and items respectively represent hyperedges and vertices. Essential and disjunctive closed itemsets can also be considered as specific cases of error-tolerant itemsets [58].

Our work can also be linked with that recently proposed in [48]. This latter work presents a general framework for setting closure operator associated to some measures through the introduction of the so-called condensed function. In comparison to our work, that of [48] does not propose any concise representation for frequent itemsets using condensable functions. In addition, the authors did not pay attention to the corresponding link between the power-set of items and that of transactions as we did in our work. On the other hand, the authors in [48], neither stated explicitly that, once an anti-monotone constraint applied, the obtained set of closed itemsets – adequate to a condensable function \( f \) (e.g. \( \mathcal{RC} \) for the disjunctive support as measure and \( \text{minsupp} \) as constraint) – may not be an exact representation of itemsets adequate to \( f \), nor highlighted an error bound or the need for adding other elements, as we did here, ensuring the exactness of the regeneration process.

In [61], the authors proposed connection operators to link \( \mathcal{P}(\mathcal{F}) \) and \( \mathcal{P}(\mathcal{T}) \) for the case of disjunctive Boolean expressions, called OR-clauses. Nevertheless, their definition of the operator \( \psi \) linking \( \mathcal{P}(\mathcal{F}) \) to \( \mathcal{P}(\mathcal{T}) \) depends on the operator \( \phi \) ensuring the dual direction and was not independently given. In addition, the authors neither gave the expression of the resulting closure operator nor carried out a thorough analysis of inherent theoretical properties. Finally, it is important to mention that these connection operators were used in [61] in order to extract closed OR-clauses (the equivalent of disjunctive closed itemsets in our case) whose disjunctive supports are encompassed between two user-defined thresholds. However, they were not applied within the framework of concise representations for frequent itemsets.

The disjunctive connector has also been used to define exact concise representations of frequent itemsets, like those based on disjunction-free sets [10,11] and (generalized) disjunction-free generators [28]. These representations only explore the conjunctive search space and required the introduction of what is called a disjunctive rule [10]. Note that we focused in this work on exact concise representations of frequent itemsets, although many approximate concise representations were proposed in the literature, like maximal frequent itemsets [6] (and their dual, i.e., minimal infrequent itemsets), \( \delta \)-free sets [9], condensed frequent pattern bases [41], \( \delta \)-clusters [56], and \( \delta \)-tolerance closed frequent itemsets [18]. Although they offer very high compactness rates, we did not compare our work to these representations since they do not allow deriving the exact frequency of itemsets. Moreover, their accuracy closely depends on the tolerated error bound.

The representation compactness is the main criterion that is generally used to compare concise representations. Other criteria such as algorithm running time or semantic of representation can also be used. Using running time relies on algorithm implementation which is our further goal. From the semantic aspect, our representation is more richer since it conveys both disjunctive and conjunctive supports of itemsets while the others rely only on the conjunctive support. Contrary to frequent closed itemsets, disjunctive closed ones offer the possibility to take into account complementary information, i.e., items that are for example mutually exclusive. For a given itemset \( I \), its disjunctive closure gathers items whose appearances depend on that of a nonempty subset of \( I \). This is not possible using frequent closed itemsets since this requires that all items of \( I \) simultaneously appear. Moreover, for an arbitrary itemset \( I \), its associated closed itemset only gives an idea about the set of items \( S \) that closely depend on all the items of \( I \). However, an item \( i \in S \) can appear in a transaction that does not contain \( I \), but only a proper subset. On the other hand, the disjunctive closure of \( I \) gathers items that closely depend on the set of items contained in \( I \). Indeed, the membership of an item to the disjunctive closure of \( I \) requires that a subset of \( I \) appears in the transactions. This can for example be useful for analyzing gene-expression data through localizing groups of genes of which the appearance depends on other groups. In addition, disjunctive closed itemsets offer an interesting starting point for the extraction of generalized association rules [39,52] which can be useful in some real-life applications. Indeed, in addition to the conjunctive support, it offers direct access to the disjunctive support of frequent itemsets, and hence to their negative support through De Morgan’s law.

8. Conclusion and perspectives

In this work, we introduced a new disjunctive closure operator and we thoroughly studied its theoretical properties. Based on this operator, we structurally characterize the disjunctive search space. Then, we introduced a new concise representation of frequent itemsets based on the disjunctive closed itemsets having at least a frequent essential itemset as a seed. In addition to interesting compactness rates, this representation allows a straightforward computation of the disjunctive and negative supports. Moreover, it is only composed of disjunctive closed itemsets which ensure its homogeneity. In nearly all experiments we performed, the obtained results showed that our representation is significantly smaller than the pioneering representations of the literature. Therefore we have proposed a concise representation (model) of frequent itemsets that distill the meaningful information with respect to the MDL principle, especially in the case of dense datasets.

Other avenues for future work mainly address the following points. We intend to address as a next step a thorough analysis of both computational time and memory consumption required for mining our representation vs. those of the literature.
and, then, deriving the whole set of frequent itemsets. In this respect, efficient algorithms for mining conjunctive closed itemsets [like DCL-CLOSED [33] and LCM [55]] could be adapted to mining disjunctive closed itemsets. Indeed, we have established the relation between disjunctive and conjunctive closed itemsets. Our investigations show that this important issue is highly correlated with that of determining the relation between their associated closure operators applied, respectively, on a given dataset and its dual. This latter dataset is obtained by replacing the presence of an item in the initial dataset by its absence and vice-versa. This issue will make it possible setting up a hybrid approach aiming at exploring either the conjunctive or the disjunctive search space according to dataset characteristics. The detailed study of this issue is currently under investigation. Another important task consists in overcoming the lack of semantics studies related to concise representations as well as helping end-users to choose the most appropriate concise representation that suits their needs according to specified constraints [8] or interestingness measures [23]. The study of the possible extension of our representation to other pattern classes [2,36,43,57] should also be investigated. Finally, the application of our representation within the framework of generalized association rules will be thoroughly addressed. Indeed, setting up a theoretical framework that includes different kinds of operators is of paramount importance for going beyond classic association rules [49]. Such an exploration can exploit the results offered by the general GUHA approach [24].

Acknowledgements

We would like to thank the anonymous reviewers for their helpful comments and suggestions. We are also grateful to T. Calders and B. Goethals for providing the source code allowing the extraction of frequent non-derivable itemsets, A. Casali and S. Lopes for providing the source code allowing the extraction of the concise representation based on frequent essential itemsets, J. Muhonen and H. Toivonen for providing the source code allowing the extraction of frequent closed non-derivable itemsets, and T. Uno for providing the source code allowing the extraction of frequent (closed) itemsets. We also express our gratitude to J. Old and J.J. Givry for English proofreading. This work was partially supported by the French-Tunisian project CMCU-Utiqte 05G1412.

Appendix: Proof of Proposition 2

\(-\text{Property (1)}\) \(T_1 \subseteq T_2 \Rightarrow f(T_1) \subseteq f(T_2)\).

- **Suppose** that \(T_1 \subseteq T_2\). If \(i \in f(T_1)\), then from \text{Definition 5}, we have \((\exists t \in T_1)((t,i) \in \mathcal{R} \land (\forall t \in T \setminus T_1)((t,i) \notin \mathcal{R}))\).

Since by hypothesis \(T_1 \subseteq T_2\), then \((\exists t \in T_2)((t,i) \in \mathcal{R})\). Let us show that \(i\) verifies the second clause. Since we have \((\forall t \in T \setminus T_1)((t,i) \notin \mathcal{R})\), then \((\forall t \in T \setminus T_2)((t,i) \notin \mathcal{R})\) also holds. Hence, \((\exists t \in T_2)((t,i) \in \mathcal{R} \land (\forall t \in T \setminus T_2)((t,i) \notin \mathcal{R}))\) is true. This implies that \(i \in f(T_2)\). We conclude that \(f(T_1) \subseteq f(T_2)\).

- **Property (2)** \(f(\mathcal{R} \cup \mathcal{S}) \subseteq f(\mathcal{R})\).

- **Let** \(i \in \mathcal{I}\). By definition (cf. \text{Definition 6}), we have \(f(\mathcal{R} \cup \mathcal{S}) = \{i \in \mathcal{I} \mid (\forall t \in T \setminus T_1)((t,i) \in \mathcal{R} \land (\forall t \in T \setminus T_2)((t,i) \notin \mathcal{R}))\} \Rightarrow (\exists i \in \mathcal{I} \mid (t,i) \in \mathcal{R}\). If we take \(i = i\), then \(i \in f(\mathcal{R} \cup \mathcal{S})\). We then conclude that \(i \subseteq f(\mathcal{R} \cup \mathcal{S})\).

- **Property (3)** \((\mathcal{R} \cup \mathcal{S}) \subseteq f(\mathcal{T})\).

- **Let** \(t \in f(\mathcal{T})\). According to the definition of \(f(\mathcal{T})\) (cf. \text{Definition 6}), we deduce that \(t\) verifies \((\exists i \in \mathcal{I} \mid (t,i) \in \mathcal{R}) \land (\forall t \in T \setminus T_1)((t,i) \notin \mathcal{R})\). We then conclude that \(t \in f(\mathcal{T})\).

- **Property (4)** \(f(\mathcal{T}) \subseteq f(\mathcal{T})\).

- **We** will prove this property by proving the inclusion in both directions.

\(-\text{Proof} (\subseteq)\)

We have \(g(T) \subseteq T\) (according to \text{Property (2)}). Hence, \(f \circ g(T) \subseteq f(T)\) (according to \text{Property (1)}).

\(-\text{Proof} (\supseteq)\)

We have \(I \subseteq f(T)\) (according to \text{Property (2)}). For the particular case where \(I = f(T)\) and by replacing \(I\) by \(f(T)\), we obtain \(g(T) \subseteq f(\mathcal{T})\).

We can then conclude that \(f(T) = f \circ g(T)\).

- **Property (4)** \(g(T) = g \circ f(T)\).

Please cite this article in press as: T. Hamrouni et al., Sweeping the disjunctive search space towards mining new exact concise representations of frequent itemsets, Data Knowl. Eng. (2009), doi:10.1016/j.datak.2009.05.001
We will prove this property by proving the inclusion in both directions.

(\subseteq)

We have \(I \subseteq f \circ g(I)\) (according to Property (2)). Hence, \(g(I) \subseteq g \circ f \circ g(I)\) (according to Property (1')).

(2)

We have \(g \circ f(T) \subseteq T\) (according to Property (2')). In particular, for \(T = g(I)\) and by replacing \(T\) by \(g(I)\), we obtain \(g \circ f \circ g(I) \subseteq g(I)\).

We can then conclude that \(g(I) = g \circ f \circ g(I)\).

- Property (5) \((f \circ g)\) is idempotent

\(f \circ g = g \circ f \circ g = g\).

- Property (4') We have \(g(I) = g \circ f \circ g(I)\) (according to Property (4')). By applying \(f\) on both sides of the equality, we obtain \(f \circ g(I) = f \circ g \circ f \circ g(I)\).

- Property (4) \((g \circ f)\) is idempotent

\(g \circ f = f \circ g \circ f = f\).

- Property (5) \((g \circ f)\) is idempotent

\(g \circ f = f \circ g \circ f = f\).

- Property (6) \(g(I) \subseteq T \iff I \subseteq f(T)\)

We will prove this equivalence by proving that both implications hold.

(\Rightarrow)

Suppose that \(g(I) \subseteq T\). Then, we have \(f \circ g(I) \subseteq f(T)\) (according to Property (1)). Since we also have \(I \subseteq f \circ g(I)\) (according to Property (2)), we conclude by transitivity that \(I \subseteq f(T)\).

(\Leftarrow)

Suppose that \(I \subseteq f(T)\). Then, we have \(g(I) \subseteq g \circ f(T)\) (according to Property (1')). Since we also have \(g \circ f(T) \subseteq T\) (according to Property (2')), we conclude by transitivity that \(g(I) \subseteq T\).

Hence, \(g(I) \subseteq I \iff f(T)\).

References


Tarek Hamrouni is a Ph.D. student at both University El-Manar (Faculty of sciences, Tunis, Tunisia) and University of Artois (Lens, France), since 2006. He obtained his master thesis in July 2005 from the Faculty of Sciences of Tunis. Since, he is a teaching contract employee at the Computer Sciences Department of the Faculty of Sciences (Tunis). His research interests include efficient extraction of informative and compact covers of pattern sets and visualization of association rules. He has been technical reviewer of several international conferences, as ICFCA, CLA and FUZZ-IEEE. He has published several papers in international conferences and journals.
Sadok Ben Yahia is an Associate Professor at the Computer Sciences department of the Faculty of Sciences of Tunis since October 2002. He obtained his Ph.D. in Computer Sciences from the Faculty of Sciences of Tunis in September 2001. Currently, he is leading a small group of researchers in Tunis, whose research interests include efficient extraction of informative and compact covers of association rules, visualization of association rules and soft computing. He has served in the program committees of over eight international conferences and has been technical reviewer of different international journals. He has also served as co-guest editor of three international journals special issues. He is currently member of the steering committee of the international conference on Concept Lattices and their Applications (CLA).

Engelbert Mephu Nguifo is currently a full professor of computer science at the Université Blaise-Pascal (Clermont-Ferrand II) and is member of the laboratory of Computer Science and Systems Modelisation and Optimization (LIMOS, CNRS UMR 6158) since 2008. Mephu Nguifo has a Ph.D. in computer science from the University of Montpellier (France). He was assistant and associate professor at University of Artois (Lens, France) where he has served as head of Computer Science department, and head of the technology transfer service, of the Institute of Technology (Lens). He has authored and co-authored more than eighty technical papers and book chapters. He has served in the program committees of over 10 international conferences and has been technical reviewer of different international journals. He has also served as guest editors of six international journals. He is currently member of the steering committee of the international conference on Concept Lattices and their Applications (CLA). His research interests include data mining, machine learning, bioinformatics, human–computer interaction, information visualization, knowledge representation, and formal concept analysis. He is a member of the ACM and ACM-SIGKDD, and French AI, bioinformatics, data mining, and classification associations (AFIA, SFBI, EGC, and SFC).