Bridging Conjunctive and Disjunctive Search Spaces for Mining a New Concise and Exact Representation of Correlated Patterns

Nassima Ben Younes, Tarek Hamrouni, and Sadok Ben Yahia

URPAH, Computer Science Department, Faculty of Sciences of Tunis, Tunis, Tunisia
benyounes.nassima@gmail.com,
{tarek.hamrouni,sadok.benyahia}@fst.rnu.tn

Abstract. In the literature, many works were interested in mining frequent patterns. Unfortunately, these patterns do not offer the whole information about the correlation rate amongst the items that constitute a given pattern since they are mainly interested in appearance frequency. In this situation, many correlation measures have been proposed in order to convey information on the dependencies within sets of items. In this work, we adopt the correlation measure *bond*, which provides several interesting properties. Motivated by the fact that the number of correlated patterns is often huge while many of them are redundant, we propose a new exact concise representation of frequent correlated patterns associated to this measure, through the definition of a new closure operator. The proposed representation allows not only to efficiently derive the correlation rate of a given pattern but also to exactly offer its conjunctive, disjunctive and negative supports. To prove the utility of our approach, we undertake an empirical study on several benchmark data sets that are commonly used within the data mining community.

Keywords: Concise representation, Correlated pattern, *bond* measure, Closure operator, Equivalence class, Conjunctive support, Disjunctive support.

1 Introduction and Motivations

In data mining, frequent pattern mining from a data set constitutes an important step within the overall knowledge extraction process. Since its inception, this key task grasped the interest of many researchers since frequent patterns constitute a source of information on the relations between items. Unfortunately, the number of mined patterns from a real-life database is often huge. As a consequence, many concise representations of frequent patterns appeared in the literature. These representations are associated to different quality measures. However, the most used one is the frequency measure (*aka* the conjunctive support or, simply, support) since it sheds light on the simultaneous appearances of items in the data set. Beyond this latter measure, recently some works [8,13] have taken into account another measure, called *disjunctive support*, which conveys information about the complementary occurrences of items. However, the size of these representations remains voluminous and many frequent patterns, having weakly correlated items, are often extracted. Moreover, whenever the minimum support threshold,

B. Pfahringer, G. Holmes, and A. Hoffmann (Eds.): DS 2010, LNAI 6332, pp. 189–204, 2010.

denoted *minsupp*, is set very low, a huge number of patterns will be generated. Additionally, within the mined set of patterns, a large portion of them is redundant or uninformative. In this situation, setting a high value of *minsupp* can solve this problem, however many interesting patterns will be missed. Therefore, in order to overcome this problem and to reduce the size of representations, many correlation measures were proposed in the literature [11,12,15,18,24]. The mined correlated patterns have then been proven to be interesting for various application domains, such as text mining, bioinformatics, market basket study, and medical data analysis, etc.

To choose the appropriate measure w.r.t. a specific aim, there are various criteria which help the user in his choice. In our case, we are interested in the *bond* measure [18]. Indeed, in addition to the information on items correlations conveyed by this measure, it offers valuable information about the conjunctive support of a pattern as well as its disjunctive and negative supports. In spite of its advantages that can be exploited in several application contexts, few studies were dedicated to the *bond* measure. One of the main reasons of this negligence is that the extraction of correlated patterns w.r.t. *bond*, is proved to be more difficult than that of correlated patterns associated to other measures, like *all-confidence*, as mentioned in [15]. In this paper, we will study the behavior of the *bond* measure w.r.t. some key criteria. We then introduce a new exact concise representation of frequent correlated patterns associated to this measure. This representation – based on a new closure operator – relies on a simultaneous exploration of both conjunctive and disjunctive search spaces, whose associated patterns are respectively characterized through the conjunctive and disjunctive supports. Indeed, in a rough manner, this new representation can be considered as a compromise between both exact representations based, respectively, on the frequent closed patterns [19] and the disjunctive closed patterns [8]. Thus, it also offers the main complementary advantages of these representations, such as the direct derivation of the conjunctive and disjunctive supports of a given pattern. Interestingly enough, the proposed representation makes it also possible to find the correlation dependencies between items of a given data set without the costly computation of the inclusion-exclusion identities [5]. To the best of our knowledge, this representation is the first one proposed in the literature associated to the *bond* measure.

The remainder of the paper is organized as follows: Section 2 presents the background used throughout the paper. We also discuss related work in Section 3. Section 4 details the f_{bond} closure operator and its main properties. Moreover, it presents the new concise representation of frequent correlated patterns associated to the *bond* correlation measure. The empirical evidences about the utility of our representation are provided in Section 5. The paper ends with a conclusion of our contributions and sketches forthcoming issues in Section 6.

2 Key Notions

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

Definition 1. - ***Data set*** - *A data set is a triplet* $\mathcal{D} = (\mathcal{T}, \mathcal{I}, \mathcal{R})$ *where* $\mathcal{T}$ *and* $\mathcal{I}$ *are, respectively, a finite set of transactions and items, and* $\mathcal{R} \subseteq \mathcal{T} \times \mathcal{I}$ *is a binary relation between the transaction set and the item set. A couple* $(t, i) \in \mathcal{R}$ *denotes that the transaction* $t \in \mathcal{T}$ *contains the item* $i \in \mathcal{I}$.

Example 1. In the remainder, we will consider the running data set $\mathcal{D}$ given in Table 1.

Table 1. An example of a data set

	A	B	C	D	E	F
1			×	×		×
2	×	×		×		
3	×	×	×	×	×	×
4	×	×	×		×	
5			×	×		×

A pattern can be characterized by three kinds of supports presented by Definition 2.

Definition 2. - ***Supports of a pattern*** - *Let $\mathcal{D} = (\mathcal{T}, \mathcal{I}, \mathcal{R})$ be a data set and I be a non-empty pattern. We mainly distinguish three kinds of supports related to I:*
- Conjunctive support: $Supp(\wedge I) = |\{t \in \mathcal{T} \mid (\forall\, i \in I, (t,i) \in \mathcal{R})\}|$
- Disjunctive support: $Supp(\vee I) = |\{t \in \mathcal{T} \mid (\exists\, i \in I, (t,i) \in \mathcal{R})\}|$
- Negative support: $Supp(\neg I) = |\{t \in \mathcal{T} \mid (\forall\, i \in I, (t,i) \notin \mathcal{R})\}|$

Example 2. Let us consider the data set of Table 1. We have $Supp(\wedge(BE)) = |\{3,4\}| = 2$. [1] $Supp(\vee(BE)) = |\{2,3,4\}| = 3$. Moreover, $Supp(\neg(BE)) = |\{1,5\}| = 2$.

Note that $Supp(\wedge\emptyset) = |\mathcal{T}|$ since the empty set is included in all transactions, while $Supp(\vee\emptyset) = 0$ since the empty set does not contain any item [13]. Moreover, $\forall\, i \in \mathcal{I}$, $Supp(\wedge i) = Supp(\vee i)$, while in the general case, for $I \subseteq \mathcal{I}$ and $I \neq \emptyset$, $Supp(\wedge I) \leq Supp(\vee I)$. A pattern I is said to be *frequent* if $Supp(\wedge I)$ is greater than or equal to a user-specified minimum support threshold, denoted *minsupp* [1]. The following lemma shows the links that exist between the different supports of a non-empty pattern I. These links are based on the *inclusion-exclusion identities* [5].

Lemma 1. - ***Inclusion-exclusion identities*** - *The inclusion-exclusion identities ensure the links between the conjunctive, disjunctive and negative supports of a non-empty pattern I.*

$$Supp(\wedge I) = \sum_{\emptyset \subset I_1 \subseteq I} (-1)^{|I_1| - 1}\, Supp(\vee I_1) \qquad (1)$$

$$Supp(\vee I) = \sum_{\emptyset \subset I_1 \subseteq I} (-1)^{|I_1| - 1}\, Supp(\wedge I_1) \qquad (2)$$

$$Supp(\neg I) = |\mathcal{T}| - Supp(\vee I)\ (\textit{The De Morgan's law}) \qquad (3)$$

An operator is said to be a closure operator if it is *extensive*, *isotone* and *idempotent* [6]. We present patterns that help to delimit the equivalence classes induced by the conjunctive closure operator f_c [19] and the disjunctive closure operator f_d [8], respectively.

Definition 3. *[19] -* ***Conjunctive closure of a pattern*** *- The conjunctive closure of a pattern $I \subseteq \mathcal{I}$ is: $f_c(I) = \max_{\subseteq}\{I' \subseteq \mathcal{I} \mid (I \subseteq I')$ and $(Supp(\wedge I') = Supp(\wedge I))\} = I \cup \{i \in \mathcal{I} \backslash I \mid Supp(\wedge I) = Supp(\wedge (I \cup \{i\}))\}$.*

[1] We use a separator-free form for the sets, *e.g.*, BE stands for the set of items {B, E}.

A minimal element within a conjunctive equivalence class is called *minimal generator* and is defined as follows.

Definition 4. *[19] - **Minimal generator** - A pattern $I \subseteq \mathcal{I}$ is said to be minimal generator if and only if Supp($\wedge I$) $< \min\{Supp(\wedge I \backslash \{i\}) \mid i \in I\}$.*

The following definition formally introduces a disjunctive closed pattern.

Definition 5. *[8] - **Disjunctive closure of a pattern** - The disjunctive closure of a pattern $I \subseteq \mathcal{I}$ is: $f_d(I) = \max_{\subseteq}\{I_1 \subseteq \mathcal{I} | (I \subseteq I_1) \wedge (Supp(\vee I) = Supp(\vee I_1))\} = I \cup \{i \in \mathcal{I} \backslash I |\ Supp(\vee I) = Supp(\vee (I \cup \{i\}))\}$.*

The antipode of a disjunctive closed pattern within the associated disjunctive equivalence class is called *essential pattern* and is defined as follows.

Definition 6. *[4] - **Essential pattern** - A pattern $I \subseteq \mathcal{I}$ is said to be essential if and only if Supp($\vee I$) $> \max\{Supp(\vee I \backslash \{i\}) \mid i \in I\}$.*

Definition 7 and Definition 8 introduce some properties that are interesting for the evaluation of quality measures, while Definition 9 and Definition 10 describe interesting pruning strategies that will be used in the remainder for reducing the number of generated patterns.

Definition 7. *[16] - **Descriptive or statistical measure** - A measure is said to be descriptive if its value is invariant w.r.t. the total number of transactions. Otherwise, it is said to be a statistical measure.*

Definition 8. *[21] - **Symmetric measure** - A measure μ is said to be symmetric if $\forall\ X, Y \subseteq \mathcal{I}$, $\mu(XY) = \mu(YX)$.*

Definition 9. *[17] - **Anti-monotone constraint** - Let $I \subseteq \mathcal{I}$. A constraint Q is said to be anti-monotone if $\forall I_1 \subseteq I$: I satisfies Q implies that I_1 satisfies Q.*

Definition 10. *[24] - **Cross-support patterns** - Given a threshold $t \in]0, 1[$, a pattern $I \subseteq \mathcal{I}$ is a cross-support pattern w.r.t. t if I contains two items x and y such that $\frac{Supp(\wedge x)}{Supp(\wedge y)} < t$.*

3 Related Work

Several works in the literature mainly paid attention to the extraction of frequent patterns. Nevertheless, the conjunctive support, used to estimate their respective frequency, only conveys information on items co-occurrences. Thus, it is not enough for giving the information about other kinds of items relations like their complementary occurrences as well as their mutual dependencies and inherent correlations. In order to convey information on the dependencies within sets of items and, then, to overcome the limits of the use only of the frequency measure, many correlation and similarity measures have been proposed. These latter measures were then applied in different fields like statistics, information retrieval, and data mining, for analyzing the relationships among items. For

example, *lift* and χ^2 are typical correlation measures used for mining association rules [3], while *any-confidence*, *all-confidence* and *bond* [18] are used in pattern mining to assess the relationships within sets of patterns. There are also many other interestingness measures and metrics studied and used in a variety of fields and applications in order to select the most interesting patterns w.r.t. a given task. In order to select the right measure for a given application, several key properties should be examined. Recent studies have identified a critical property, null-invariance, for measuring associations among items in large data sets, but many measures do not have this property. Indeed, in [23], the authors re-examine a set of null-invariant, *i.e.*, uninfluenced by the number of null transactions, interestingness measures and they express them as a generalized mathematical mean. However in their work, the authors only considered the application of the studied measures only for patterns of size two. Moreover, other studies are based on the analysis of measures w.r.t. some desirable properties, such as the nice property of anti-monotonicity, like carried out in [14]. In this respect, anti-monotone measures are extensively used to develop efficient algorithms for mining correlated patterns [12,15,18,24]. However, almost all dedicated works to correlated patterns do not address the problem of the huge number of mined patterns while many of them are redundant. To the best of our knowledge, only the work proposed in [12] allows the extraction of a concise representation of frequent correlated patterns based on the *all-confidence* measure. Furthermore, the proposed works only rely on the exploration of the conjunctive search space for the extraction of the correlated patterns and no one was interested in the exploration of the disjunctive search space.

In addition, our work can also be linked with that proposed in [20]. This latter work presents a general framework for setting closure operators associated to some measures through the introduction of the so-called *condensable function*. In comparison to our work, that of [20] does not propose any concise representation for frequent correlated patterns using the condensable measure *bond*. In addition, the authors neither studied the structural properties of this measure nor paid attention to the corresponding link between the patterns associated to this measure and those characterizing the conjunctive search space and the disjunctive one. All these points are addressed in the following.

4 New Concise and Exact Representation of Frequent Correlated Patterns

We concentrate now on the proposed representation of frequent correlated patterns. We firstly introduce a structural characterization of the *bond* measure and, then, detail the associated closure operator on which the representation is based.

4.1 Structural Characterization of the *bond* Measure

We study, in this subsection, different interesting properties of the *bond* measure. In the literature, other equivalent measures to *bond* are used in different application contexts such as *Coherence* [15], *Tanimoto coefficient* [22], and *Jaccard* [10]. With regard to data mining, the *bond* measure is similar to the conjunctive support but w.r.t. a subset of the data rather than the entire data set. Indeed, semantically speaking, this measure conveys

the information about the correlation of a pattern I by computing the ratio between the number of co-occurrences of its items and the cardinality of its universe, which is equal to the transaction set containing a non-empty subset of I. It is worth mentioning that, in the previous works dedicated to this measure, the disjunctive support has never been used to express it. Indeed, none of these works highlighted the link between the denominator – the cardinality of the universe of I – and the disjunctive support. Thus, we propose a new expression of *bond* in Definition 11.

Definition 11. - ***The bond measure*** - *The bond measure of a non-empty pattern $I \subseteq \mathcal{I}$ is defined as follows:*

$$bond(I) = \frac{Supp(\wedge I)}{Supp(\vee I)}$$

The use of the disjunctive support allows to reformulate the expression of the *bond* measure in order to bring out some pruning conditions for the extraction of the patterns fulfilling this measure. Indeed, as shown later, the *bond* measure satisfies several properties that offer interesting pruning strategies allowing to reduce the number of generated pattern during the extraction process. Note that the value of the *bond* measure of the empty set is undefined since its disjunctive support is equal to 0. However, this value is positive because $\lim_{I \mapsto \emptyset} bond\ (I) = \frac{|\mathcal{T}|}{0} = +\infty$. As a result, the empty set will be considered as a correlated pattern for any minimal threshold of the *bond* correlation measure. To the best of our knowledge, none of the literature works was interested in the properties of this measure in the case of the empty set.

The following proposition presents interesting properties verified by *bond*.

Proposition 1. - ***Some properties of the bond measure*** - *The bond measure is descriptive and symmetric.*

Proof. The numerator of the bond measure represents the conjunctive support of a pattern I, while the denominator represents its disjunctive support. Being the ratio between two descriptive and symmetric measures, bond is also descriptive and symmetric.

Several studies [21,23] have shown that it is desirable to select a descriptive measure that is not influenced by the number of transactions that contain none of pattern items. The symmetric property fulfilled by the *bond* measure makes it possible not to treat all the combinations induced by the precedence order of items within a given pattern. Noteworthily, the anti-monotony property, fulfilled by the *bond* measure as proven in [18], is very interesting. Indeed, all the subsets of a correlated pattern are also necessarily correlated. Then, we can deduce that any pattern having at least one uncorrelated proper subset is necessarily uncorrelated. It will thus be pruned without computing the value of its *bond* measure. In the next proposition, we introduce the relationship between the *bond* measure and the cross-support property.

Proposition 2. - ***Cross-support property of the bond measure*** - *Any cross-support pattern $I \subseteq \mathcal{I}$, w.r.t. a threshold $t \in]0,1[$, is guaranteed to have bond(I)$< t$.*

Proof. Let $I \subseteq \mathcal{I}$ and $t \in]0,1[$. If I is a cross-support pattern w.r.t. the threshold t, then $\exists\ x$ and $y \in I$ such as $\frac{Supp(\wedge x)}{Supp(\wedge y)} < t$. Let us prove that bond(I) $< t$: bond(I) $= \frac{Supp(\wedge (I))}{Supp(\vee (I))} \leq \frac{Supp(\wedge (xy))}{Supp(\vee (xy))} \leq \frac{Supp(\wedge (xy))}{Supp(\vee y)} \leq \frac{Supp(\wedge x)}{Supp(\vee y)} = \frac{Supp(\wedge x)}{Supp(\wedge y)} < t$.

The cross-support property is very important. Indeed, any pattern, containing two items fulfilling the cross-support property w.r.t. a minimal threshold of correlation, is not correlated. Thus, this property avoids the computation of its conjunctive and disjunctive supports, required to evaluate its *bond* value.

The set of frequent correlated patterns associated to *bond* is defined as follows.

Definition 12. - ***The set of frequent correlated patterns*** - *Considering the support threshold minsupp and the correlation threshold minbond, the set of frequent correlated patterns, denoted $\mathcal{FCP}$, is equal to: $\mathcal{FCP} = \{I \subseteq \mathcal{I} \mid bond(I) \geq minbond$ and $Supp(\wedge I) \geq minsupp\}$.*

The following proposition establishes the relation between the values of the *bond* measure as well as the conjunctive and disjunctive supports of two patterns linked by set inclusion.

Proposition 3. *Let I and I_1 be two patterns such as $I \subseteq I_1 \subseteq \mathcal{I}$. We have $bond(I) = bond(I_1)$ if and only if $Supp(\wedge I) = Supp(\wedge I_1)$ and $Supp(\vee I) = Supp(\vee I_1)$.*

Proof. The bond correlation measure of a pattern is the ratio between its conjunctive and disjunctive supports. So, if there is two patterns I and $I_1 \subseteq \mathcal{I}$, with $I \subseteq I_1$, and if they have equal values of the bond measure, they also have equal values of the conjunctive and disjunctive supports. Indeed, to have $\frac{a}{b} = \frac{c}{d}$ (where a, b, c, and d are four positive integers), three cases are possible: ($a = c$ and $b = d$) or ($a > c$ and $b > d$) or ($a < c$ and $b < d$), such that $a \times d = b \times c$. So, when we add an item i to the pattern I, its conjunctive and disjunctive supports vary inversely proportionally to each other such that $\forall i \in \mathcal{I}$, $Supp(\wedge I) \geq Supp(\wedge(I \cup \{i\}))$ and $Supp(\vee I) \leq Supp(\vee(I \cup \{i\}))$. Thus, the unique possibility to have $\frac{Supp(\wedge I)}{Supp(\vee I)} = \frac{Supp(\wedge(I \cup \{i\}))}{Supp(\vee(I \cup \{i\}))}$ occurs when $Supp(\wedge I) = Supp(\wedge(I \cup \{i\}))$ and $Supp(\vee I) = Supp(\vee(I \cup \{i\}))$. In an incremental manner, it can be easily shown that whenever $\frac{Supp(\wedge I)}{Supp(\vee I)} = \frac{Supp(\wedge(I \cup I_1))}{Supp(\vee(I \cup I_1))}$, $Supp(\wedge I) = Supp(\wedge(I \cup I_1))$ and $Supp(\vee I) = Supp(\vee(I \cup I_1))$.

4.2 Closure Operator Associated to the *bond* Measure

Since many correlated patterns share exactly the same characteristics, an interesting solution in order to reduce the number of mined patterns is to find a closure operator associated to the *bond* measure. Indeed, thanks to the non-injectivity property of the closure operator, correlated patterns having common characteristics will be mapped without information loss into a single element, namely the associated closed correlated pattern. The proposed closure operator is given by the following definition.

Definition 13. - ***The f_{bond} operator*** - *Let $\mathcal{D} = (\mathcal{T}, \mathcal{I}, \mathcal{R})$ be a data set. Let f_c and f_d be, respectively, the conjunctive closure operator and the disjunctive one. Formally, the f_{bond} operator is defined as follows:*

$$\begin{array}{rcl} f_{bond} : \mathcal{P}(\mathcal{I}) & \rightarrow & \mathcal{P}(\mathcal{I}) \\ I & \mapsto & f_{bond}(I) = I \cup \{i \in \mathcal{I} \backslash I \mid bond(I) = bond(I \cup \{i\})\} \\ & & \quad = \{i \in \mathcal{I} \mid i \in f_c(I) \cap f_d(I)\} \end{array}$$

The fact that the application of f_{bond} to a given pattern is exactly equal to the intersection of both its conjunctive and disjunctive closures, associated respectively to the f_c and f_d operators (*cf.* Definition 3 and Definition 5), results from Proposition 3.

Example 3. Consider the data set illustrated by Table 1. We have: since $f_c(AB) = AB$ and $f_d(AB) = ABE$, then $f_{bond}(AB) = AB$. Since $f_c(AC) = ABCE$ and $f_d(AC) = ABCDEF$, then $f_{bond}(AC) = ABCE$.

The next proposition proves that f_{bond} is a closure operator.

Proposition 4. *The f_{bond} operator is a closure operator.*

Proof. Let I, $I' \subseteq \mathcal{I}$ be two patterns. $f_{bond}(I) = f_c(I) \cap f_d(I)$ and $f_{bond}(I') = f_c(I') \cap f_d(I')$. Let us prove that the f_{bond} operator is a closure operator.

(1) *Extensivity*: *Let us prove that $I \subseteq f_{bond}(I)$*

$$\begin{cases} f_c \text{ is a closure operator} \Rightarrow I \subseteq f_c(I) \\ f_d \text{ is a closure operator} \Rightarrow I \subseteq f_d(I) \end{cases} \Rightarrow \{\ I \subseteq (f_c(I) \cap f_d(I)) \Rightarrow \{\ I \subseteq f_{bond}(I).$$

Thus, the f_{bond} operator is extensive.

(2) *Isotony*: *Let us prove that $I \subseteq I' \Rightarrow f_{bond}(I) \subseteq f_{bond}(I')$*

$$I' \subseteq I \Rightarrow \begin{cases} f_c(I') \subseteq f_c(I) \\ f_d(I') \subseteq f_d(I) \end{cases} \Rightarrow \{\ (f_c(I') \cap f_d(I')) \subseteq (f_c(I) \cap f_d(I))$$
$$\Rightarrow \{\ f_{bond}(I') \subseteq f_{bond}(I).$$

Thus, the f_{bond} operator is isotone.

(3) *Idempotency*: *Let us prove that $f_{bond}(f_{bond}(I)) = f_{bond}(I)$*

According to (1), we have $f_{bond}(I) \subseteq f_{bond}(f_{bond}(I))$. We will prove by absurdity that $f_{bond}(f_{bond}(I)) = f_{bond}(I)$.

Suppose that $f_{bond}(I) \subset f_{bond}(f_{bond}(I))$. This is equivalent to bond$(I) \neq$ bond$(f_{bond}(I))$.

However, this is impossible because bond$(f_{bond}(I))$ = bond(I) (cf. Proposition 3).

Thus, $f_{bond}(f_{bond}(I)) = f_{bond}(I)$, i.e., the f_{bond} operator is idempotent.

According to **(1)**, **(2)** *and* **(3)**, *the operator f_{bond} is a closure operator.*

Definition 14. *- **Closed pattern by** f_{bond} - Let $I \subseteq \mathcal{I}$ be a pattern. The associated closure $f_{bond}(I)$ is equal to the maximal set of items containing I and having the same value of bond as that of I.*

Example 4. Consider our running data set. We have the maximal set of items which have an equal value of the *bond* measure than *AF* is *ABCDEF*. Then, $f_{bond}(AF)$ = *ABCDEF*

The next definition introduces the set of frequent closures while Definition 16 presents the minimal patterns associated to f_{bond}.

Definition 15. *The set $\mathcal{FCCP}$ of frequent closed correlated patterns is equal to: $\mathcal{FCCP}$ = $\{\ I \in \mathcal{FCP} \mid$ bond$(I) >$ bond$(I \cup \{i\}), \forall\, i \in \mathcal{I} \backslash I\}$.*

Definition 16. *- **Frequent minimal correlated pattern** - Let $I \in \mathcal{FCP}$. The pattern I is said to be minimal if and only if $\forall\, i \in I$, $bond(I) < bond(I \backslash \{i\})$ or, equivalently, $\nexists\, I_1 \subset I$ such that $f_{bond}(I) = f_{bond}(I_1)$.*

Example 5. Consider our running data set illustrated by Table 1 for *minsupp = 1* and *minbond = 0.30*. The pattern CE is a minimal one since $bond(CE) < bond(C)$ and $bond(CE) < bond(E)$. Moreover, the pattern CE is correlated and frequent since $bond(CE) = 0.50 > 0.30$ and $Supp(\wedge(CE)) = 2 \geq 1$.

The next proposition links a minimal pattern with the key notions of minimal generator (*cf.* Definition 4) and essential pattern (*cf.* Definition 6) of the conjunctive and disjunctive search spaces, respectively.

Proposition 5. *Every minimal generator (resp. essential pattern) is a minimal pattern.*

Proof. Let $I \subseteq \mathcal{I}$ be a minimal generator (resp. essential pattern). $\forall\, i \in I$, $Supp(\wedge I) < Supp(\wedge(I \backslash \{i\}))$ and $Supp(\vee I) \geq Supp(\vee(I \backslash \{i\}))$ (resp. $Supp(\vee I) > Supp(\vee(I \backslash \{i\}))$ and $Supp(\wedge I) \leq Supp(\wedge(I \backslash \{i\}))$). Thus, in both cases, $\frac{Supp(\,\wedge\, I)}{Supp(\,\vee\, I)} < \frac{Supp(\,\wedge\, (I \setminus \{i\}))}{Supp(\,\vee\, (I \setminus \{i\}))}$. As a result, $bond(I) \neq bond(I \backslash \{i\})$ and, hence, I is a minimal pattern.

It is important to note that a minimal pattern can be neither an essential pattern nor a minimal generator. This is illustrated through the following example.

Example 6. Consider our running data set illustrated by Table 1. According to Example 5, CE is a minimal pattern, although it is neither a minimal generator (since $Supp(\,\wedge\, (CE)) = Supp(\,\wedge\, E)$) nor an essential pattern (since $Supp(\,\vee\, (CE)) = Supp(\,\vee\, C)$).

In the remainder, we will consider the empty set as a frequent minimal correlated pattern given that the values of its conjunctive support and that of its *bond* measure exceed both *minsupp* and *minbond* thresholds, respectively (the conjunctive support of the empty set is equal to $|\mathcal{T}| \geq$ *minsupp* and the value of its *bond* measure tends to $+\infty$ when I tends to $\emptyset$). Besides, we will consider the closure of the empty set as equal to itself. Let us note that these considerations are important and allow the set of frequent minimal correlated patterns to be flagged as *order ideal* (*aka downward closed set*) [6], without having any effect neither on the supports of the other patterns nor on their closures.

Proposition 6. *The set $\mathcal{FMCP}$ of the frequent minimal correlated pattern is an order ideal. Thus, it fulfills the following properties:*

- If $X \in \mathcal{FMCP}$, then $\forall\, Y \subseteq X$, $Y \in \mathcal{FMCP}$, i.e., the constraint "be a frequent minimal correlated pattern" is anti-monotone.

- If $X \notin \mathcal{FMCP}$, then $\forall\, Y \supseteq X$, $Y \notin \mathcal{FMCP}$, i.e., the constraint "not to be a frequent minimal correlated pattern" is monotone.

Proof. The proof results from the following fact: for $i \in \mathcal{I}$ and for all $X \subseteq Y \subset \mathcal{I}$, if $bond(X \cup \{i\}) = bond(X)$, then $bond(Y \cup \{i\}) = bond(Y)$, i.e., if $(X \cup \{i\})$ is not a minimal pattern, so $(Y \cup \{i\})$ is also not minimal. The constraint "not to be a minimal pattern" is hence monotone, w.r.t. set inclusion. We deduce that the constraint "be a minimal pattern" is anti-monotone. Moreover, the constraint "be a frequent minimal correlated pattern" is anti-monotone since resulting from the conjunction of three

anti-monotone constraints: "to be frequent", "to be correlated", and "to be minimal". Conversely, the constraint "not to be a frequent minimal correlated pattern" is monotone. We then deduce that the set $\mathcal{FMCP}$ is an order ideal.

The closure operator f_{bond} induces an equivalence relation on the power-set of the set of items $\mathcal{I}$, which splits it into disjoint subsets, called f_{bond} *equivalence classes*. In each class, all the elements have the same f_{bond} closure and the same value of *bond*. The minimal patterns of a *bond* equivalence class are the smallest incomparable members, w.r.t. set inclusion, while the f_{bond} closed pattern is the largest one.

To establish the link with the conjunctive and disjunctive search spaces, an f_{bond} equivalence class as well as conjunctive and disjunctive classes are given in Figure 1. The equivalence class associated to the *bond* measure can then be considered as an intermediary representation of both conjunctive and disjunctive ones.

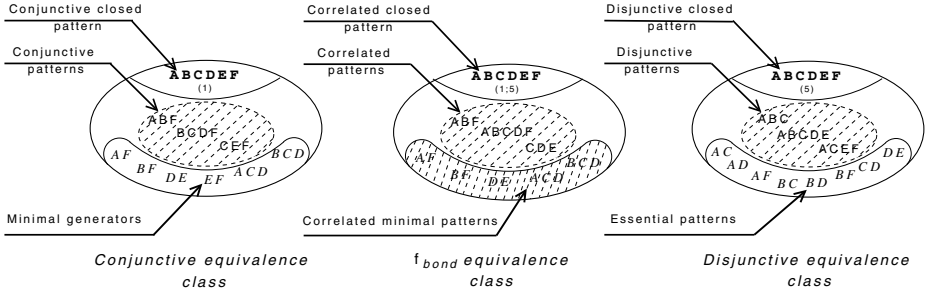

Fig. 1. Structural characterization of the equivalence classes associated respectively (from left to right) to the f_c, f_{bond}, and f_d closure operators w.r.t. the data set given in Table 1.

4.3 New Concise Representation Associated to the *bond* Measure

Based on the f_{bond} closure operator, we can design two representations which cover the same frequent correlated patterns. The first is based on the frequent closed correlated patterns, whereas the second one is based on the frequent minimal correlated patterns. In this work, we focus on the first one, since it is considered more concise thanks to the fact that a f_{bond} equivalence class always contains only one closed pattern, but potentially several minimal patterns. Let us define the new concise representation of frequent correlated patterns based on the frequent closed correlated patterns associated to the *bond* measure.

Definition 17. *The representation $\mathcal{RFCCP}$ based on the set of frequent closed correlated patterns associated to f_{bond} is defined as follows:*

$$\mathcal{RFCCP} = \{(I, Supp(\wedge I), Supp(\vee I)) \mid I \in \mathcal{FCCP}\}.$$

Example 7. Consider our running data set illustrated by Table 1. For *minsupp = 2* and *minbond = 0.60*, the representation $\mathcal{RFCCP}$ of the $\mathcal{FCP}$ set is equal to: $\{(\emptyset, 5, 0)$, (*C*, 4, 4), (*D*, 4, 4), (*E*, 2, 2), (*F*, 3, 3), (*AB*, 3, 3), (*CF*, 3, 4), (*DF*, 3, 4), (*ABE*, 2, 3), (*CDF*, 3, 5)$\}$.

The next theorem proves that the proposed representation is an exact one of frequent correlated patterns.

Theorem 1. *The representation $\mathcal{RFCCP}$ constitutes an exact concise representation of the $\mathcal{FCP}$ set.*

Proof. Thanks to a reasoning by recurrence, we will demonstrate that, for an arbitrary pattern $I \subseteq \mathcal{I}$, its f_{bond} closure, $f_{bond}(I)$, belongs to $\mathcal{FCCP}$ if it is frequent correlated. In this regard, let $\mathcal{FMCP}_k$ be the set of frequent minimal correlated patterns of size k and $\mathcal{FCCP}_k$ be the associated set of closures by f_{bond}. The hypothesis is verified for single items i inserted in $\mathcal{FMCP}_1$, and their closures $f_{bond}(i)$ are inserted in $\mathcal{FCCP}_1$ if Supp($\wedge i$) $\geq$ minsupp (since $\forall\, i \in \mathcal{I}$, bond($i$) $= 1 \geq$ minbond). Thus, $f_{bond}(i) \in \mathcal{FCCP}$. Now, suppose that $\forall I \subseteq \mathcal{I}$ such as $|I| = n$. We have $f_{bond}(I) \in \mathcal{FCCP}$ if I is frequent correlated. We show that, $\forall I \subseteq \mathcal{I}$ such as $|I| = (n+1)$, we have $f_{bond}(I) \in \mathcal{FCCP}$ if I is frequent correlated. Let I be a pattern of size $(n+1)$. Three situations are possible:
(a) *if $I \in \mathcal{FCCP}$, then necessarily $f_{bond}(I) \in \mathcal{FCCP}$ since f_{bond} is idempotent.*
(b) *if $I \in \mathcal{FMCP}_{n+1}$, then $f_{bond}(I) \in \mathcal{FCCP}_{n+1}$ and, hence, $f_{bond}(I) \in \mathcal{FCCP}$.*
(c) *if I is neither closed nor minimal – $I \notin \mathcal{FCCP}$ and $I \notin \mathcal{FMCP}_{n+1}$ – then $\exists I_1 \subset I$ such as $|I_1| = n$ and bond(I) = bond(I_1). According to Proposition 3, $f_{bond}(I) = f_{bond}(I_1)$, and I is then frequent correlated. Moreover, using the hypothesis, we have $f_{bond}(I_1) \in \mathcal{FCCP}$ and, hence, $f_{bond}(I) \in \mathcal{FCCP}$.*

It is worth noting that maintaining both conjunctive and disjunctive supports for each pattern belonging to the representation allows to avoid the cost of the evaluation of the inclusion-exclusion identities. Indeed, this evaluation can be very expensive, in particular in the case of long correlated patterns to be derived. For example, for a pattern containing 20 items, the evaluation of an inclusion-exclusion identity will involve the computation of the supports of all its non-empty subsets, *i.e.*, 2^{20} - 1 terms (*cf.* Lemma 1, page 191). Such an evaluation will be mandatory if we retain only one support and not both. It will then be carried out in order to derive the non-retained support to compute the value of the *bond* measure for each pattern. Thus, contrarily to the main concise representations of the literature, the regeneration of the whole frequent correlated patterns from the representation $\mathcal{RFCCP}$ can be carried out in a very simple and effective way. Indeed, in an equivalence class associated to the *bond* measure, patterns present the same value of this measure and consequently the same conjunctive, disjunctive and negative supports. Then, to derive the information corresponding to a frequent correlated pattern, it is enough to locate the smallest frequent closed correlated pattern which covers it and which corresponds to its closure by f_{bond}. Thus, we avoid the highly costly evaluation of the inclusion-exclusion identities.

Note however that the closure operator associated to the *bond* measure induces a strong constraint. Indeed, the f_{bond} operator gathers the patterns having the same conjunctive and disjunctive supports (*cf.* Proposition 3). Consequently, the number of patterns belonging to a given equivalence class associated to this operator is in almost all cases lower than those resulting when the conjunctive and the disjunctive closure operators are separately applied. Fortunately, the pruning based on both thresholds *minsupp* and *minbond* drastically reduces the size of our concise representation, as shown in the next section.

5 Experimental Results

In this section, our objective is to show, through extensive experiments, that our concise representation based on frequent closed correlated patterns provides interesting compactness rates compared to the whole set of frequent correlated patterns. All experiments were carried out on a PC equipped with a 2 GHz Intel processor and 4 GB of main memory, running the Linux Ubuntu 9.04 (with 2 GB of swap memory). The experiments were carried out on benchmark data sets[2].

We first show that the complete set of frequent correlated patterns ($\mathcal{FCP}$) is much bigger in comparison with both that of frequent correlated closed patterns ($\mathcal{FCCP}$) and that of frequent minimal correlated patterns ($\mathcal{FMCP}$) especially for low *minsupp* and *minbond* values. In this respect, Figure 2 presents the cardinalities of these sets when *minsupp* varies and *minbond* is fixed, while, in Figure 3, cardinalities are shown when *minbond* varies and *minsupp* is fixed.

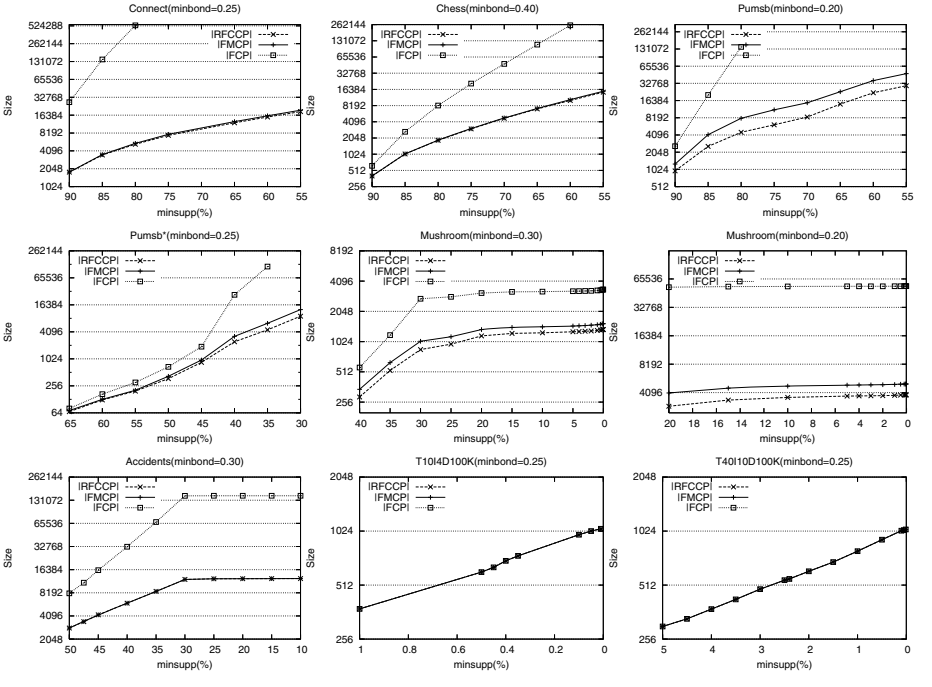

Fig. 2. Number of patterns generated when *minsupp* varies and *minbond* is fixed

The obtained results show that the size of $\mathcal{FCCP}$ is always smaller than that of $\mathcal{FMCP}$ over the entire range of the support and *bond* thresholds. For example, considering the PUMSB data set for *minsupp* = *50*% and *minbond* = *0.5*: $|\mathcal{FMCP}|$ = *68, 532*, while $|\mathcal{RFCCP}|$ = *40, 606*, with a reduction reaching approximately *41*%. These results

[2] Available at `http://fimi.cs.helsinki.fi/data`

are obtained thanks to the closure operator f_{bond} which gathers into disjoint subsets, *i.e.*, f_{bond} *equivalence classes*, patterns that have the same characteristics.

The key role of this operator is all the more visible when we compare the number of the whole set of correlated patterns with that of the proposed representation. In this respect, Figures 2 and 3 show that $\mathcal{FCCP}$ mining generates a much smaller set than that of frequent correlated patterns. Interestingly enough, compression rates increase proportionally with the decrease of the *minsupp* and *minbond* values. It is hence a desirable phenomenon since the number of frequent correlated patterns increases dramatically as far as one of both thresholds decreases. For example, let us consider the PUMSB* data set, and *minbond* fixed at *0.25%*: for *minsupp* = *60%*: $\frac{|\mathcal{FCP}|}{|\mathcal{RFCCP}|} = \frac{167}{124} = 1.34$, while for *minsupp* = *35%*: $\frac{|\mathcal{FCP}|}{|\mathcal{RFCCP}|} = \frac{116,\ 787}{4,\ 546} = 25.69 \gg 1.34$. In fact, in general, only single items can fulfill high values of thresholds. In this situation, the set of frequent correlated patterns only contains items which are in most cases equal to their closures. However, when thresholds are set very low, a high number of frequent correlated patterns, which are in general not equal to their respective closures, is extracted.

Noteworthily, the size reduction rates brought by the proposed representation, w.r.t. the size of the $\mathcal{FCP}$ set, are closely related to the chosen *minsupp* and *minbond*

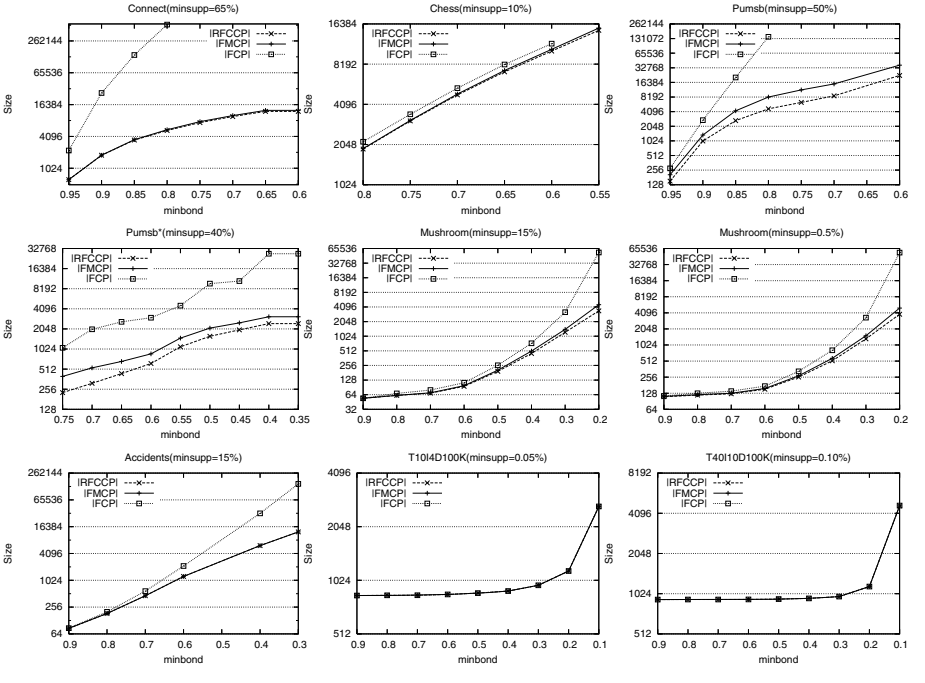

Fig. 3. Number of patterns generated when *minbond* varies and *minsupp* is fixed

values (*cf.* Figure 2 (*resp.* Figure 3) for a variation of the fixed value of *minbond* (*resp. minsupp*) for the MUSHROOM data set). Note that rates vary depending on the data set characteristics and are, hence, more important for some data sets than for others. In this respect, for CONNECT, PUMSB and PUMSB*, the obtained rates are more interesting than for CHESS, MUSHROOM and ACCIDENTS. This is explained by the fact that for the first three data sets, which contain strongly correlated items, the f_{bond} operator produces equivalence classes containing a high number of patterns. Thus, the number of closures is much more reduced compared to that of the whole set of patterns, even for high values of *minsupp* and *minbond*. While for the latter three data sets, items are relatively less correlated than for the three first ones, which decreases the number of patterns having common characteristics, and hence the same closure.

For the two data sets T10I4D100K and T40I10D100K, the size of the representation $\mathcal{RFCCP}$ is almost equal to that of the $\mathcal{FCP}$ set. This is due to the nature of these data sets, which contain a large number of items but only a few of them frequently occur. Moreover, most of them are weakly correlated with each other. This makes the size reduction rates brought by the representation meaningless in such data sets. It is important to note that, these two data sets are the "worst" for the f_{bond} closure operator as well as for the conjunctive and disjunctive closure operators (*cf.* [7,8] for experimental results associated to these two latter operators). In addition, the number of frequent correlated patterns that are extracted from these data sets is relatively reduced for each used value of *minsupp* and *minbond*.

6 Conclusion and Perspectives

In this work, we studied the behavior of the *bond* correlation measure according to some key properties. In addition, we introduced a new closure operator associated to this measure and we thoroughly studied its theoretical properties. Based on this operator, we characterized the elements of the search space associated to the *bond* measure. Then, we introduced a new concise representation of frequent patterns based on the frequent closed correlated patterns. Beyond interesting compactness rates, this representation allows a straightforward computation of the conjunctive, disjunctive and negative supports of a pattern. In nearly all experiments we performed, the obtained results showed that our representation is significantly smaller than the whole set of frequent correlated patterns.

Other avenues for future work mainly address a thorough analysis of the computational time required for mining our representation and, then, for the derivation process of the whole set of frequent patterns. In this respect, efficient algorithms for mining conjunctive closed patterns (like LCM and DCI-CLOSED [2]) and disjunctive closed patterns (like DSSRM [9]) could be adapted for mining frequent closed correlated patterns. Other important tasks consist in applying the proposed approach in real-life applications and extending it by (i) generating association rules starting from correlated frequent patterns, and, (i) extracting unfrequent (*aka* rare) correlated patterns associated to the *bond* measure by selecting the most informative ones.

References

1. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules. In: Proceedings of the 20th International Conference on VLDB 1994, Santiago, Chile, pp. 487–499 (1994)
2. Ben Yahia, S., Hamrouni, T., Mephu Nguifo, E.: Frequent closed itemset based algorithms: A thorough structural and analytical survey. ACM-SIGKDD Explorations 8(1), 93–104 (2006)
3. Brin, S., Motwani, R., Silverstein, C.: Beyond market baskets: generalizing association rules to correlations. In: Proceedings of the ACM SIGMOD International Conference on SIGMOD 1997, Tucson, Arizona, USA, pp. 265–276 (1997)
4. Casali, A., Cicchetti, R., Lakhal, L.: Essential patterns: A perfect cover of frequent patterns. In: Proceedings of the 7th International Conference on DaWaK, Copenhagen, Denmark, pp. 428–437 (2005)
5. Galambos, J., Simonelli, I.: Bonferroni-type inequalities with applications. Springer, Heidelberg (2000)
6. Ganter, B., Wille, R.: Formal Concept Analysis. Springer, Heidelberg (1999)
7. Hamrouni, T.: Mining concise representations of frequent patterns through conjunctive and disjunctive search spaces. Ph.D. Thesis, University of Tunis El Manar (Tunisia) and University of Artois (France) (2009), `http://tel.archives-ouvertes.fr/tel-00465733`
8. Hamrouni, T., Ben Yahia, S., Mephu Nguifo, E.: Sweeping the disjunctive search space towards mining new exact concise representations of frequent itemsets. Data & Knowledge Engineering 68(10), 1091–1111 (2009)
9. Hamrouni, T., Ben Yahia, S., Mephu Nguifo, E.: Optimized mining of a concise representation for frequent patterns based on disjunctions rather than conjunctions. In: Proceedings of the 23rd International Florida Artificial Intelligence Research Society Conference (FLAIRS 2010), pp. 422–427. AAAI Press, Daytona Beach, Florida, USA (2010)
10. Jaccard, P.: Nouvelles recherches sur la distribution florale. Bulletin de la Société Vaudoise des Sciences Naturelles 44, 223–270 (1908)
11. Ke, Y., Cheng, J., Yu, J.X.: Efficient discovery of frequent correlated subgraph pairs. In: Proceedings of the 9th IEEE International Conference on Data Mining, Miami, Florida, USA, pp. 239–248 (2009)
12. Kim, W.Y., Lee, Y.K., Han, J.: CCMINE: Efficient mining of confidence-closed correlated patterns. In: Proceedings of the 8th International Pacific-Asia Conference on KDD, Sydney, Australia, pp. 569–579 (2004)
13. Kryszkiewicz, M.: Compressed disjunction-free pattern representation versus essential pattern representation. In: Corchado, E., Yin, H. (eds.) IDEAL 2009. LNCS, vol. 5788, pp. 350–358. Springer, Heidelberg (2009)
14. Le Bras, Y., Lenca, P., Lallich, S.: Mining interesting rules without support requirement: a general universal existential upward closure property. Annals of Information Systems 8, 75–98 (2010)
15. Lee, Y.K., Kim, W.Y., Cai, Y.D., Han, J.: CoMine: Efficient mining of correlated patterns. In: Proceedings of the 3rd IEEE International Conference on Data Mining, Melbourne, Florida, USA, pp. 581–584 (2003)
16. Lenca, P., Vaillant, B., Meyer, P., Lallich, S.: Association rule interestingness measures: Experimental and theoretical studies. In: Quality Measures in Data Mining, Studies in Computational Intelligence, vol. 43, pp. 51–76. Springer, Heidelberg (2007)
17. Mannila, H., Toivonen, H.: Levelwise search and borders of theories in knowledge discovery. Data Mining and Knowledge Discovery 1(3), 241–258 (1997)
18. Omiecinski, E.R.: Alternative interest measures for mining associations in databases. IEEE Transactions on Knowledge and Data Engineering 15(1), 57–69 (2003)

19. Pasquier, N., Bastide, Y., Taouil, R., Stumme, G., Lakhal, L.: Generating a condensed representation for association rules. Journal of Intelligent Information Systems 24(1), 25–60 (2005)
20. Soulet, A., Crémilleux, B.: Adequate condensed representations of patterns. Data Mining and Knowledge Discovery 17(1), 94–110 (2008)
21. Tan, P.N., Kumar, V., Srivastava, J.: Selecting the right interestingness measure for association patterns. In: Proceedings of the 8th ACM SIGKDD International Conference on KDD, Edmonton, Alberta, Canada, pp. 32–41 (2002)
22. Tanimoto, T.T.: An elementary mathematical theory of classification and prediction. Technical Report, I.B.M. Corporation Report (1958)
23. Wu, T., Chen, Y., Han, J.: Re-examination of interestingness measures in pattern mining: a unified framework. Data Mining and Knowledge Discovery (2010) doi: 10.1007/s10618-009-0161-2
24. Xiong, H., Tan, P.N., Kumar, V.: Hyperclique pattern discovery. Data Mining and Knowledge Discovery 13(2), 219–242 (2006)