Motifs corrélés rares : caractérisation et nouvelles représentations concises exactes

Souad Bouasker, Tarek Hamrouni, Sadok Ben Yahia

Département des Sciences de l'Informatique, Faculté des Sciences de Tunis.
{tarek.hamrouni, sadok.benyahia}@fst.rnu.tn

Résumé. Dans la littérature, les travaux se sont principalement focalisés sur l'extraction des motifs fréquents. Toutefois, récemment, la fouille des motifs rares s'est avérée intéressante puisque ces motifs permettent de véhiculer des connaissances concernant des événements rares, inattendus. Ils ont ainsi prouvé leur grande utilité dans plusieurs domaines d'application. Cependant, un constat important associé à l'extraction des motifs rares est d'une part leur nombre très élevé et d'autre part la qualité faible de plusieurs motifs extraits. Ces derniers peuvent en effet ne pas présenter des corrélations fortes entre les items les constituant. Afin de pallier ces inconvénients, nous proposons dans cet article d'intégrer la mesure de corrélation *bond* afin d'extraire seulement l'ensemble des motifs rares vérifiant cette mesure. Une caractérisation de l'ensemble résultant, des motifs corrélés rares, est alors réalisée en se basant sur l'étude des contraintes de nature différentes induite par la rareté et la corrélation. En outre, en se basant sur les classes d'équivalence associées à un opérateur de fermeture dédié à la mesure *bond*, nous proposons des représentations concises exactes des motifs corrélés rares.

1 Introduction et motivations

L'extraction des règles d'association est une technique très répandue dans la fouille de données et répond aux besoins des experts dans plusieurs domaines d'application. Plusieurs travaux se sont ainsi focalisés sur la dérivation des règles d'association à partir des motifs fréquents. Toutefois, l'utilisation de ces motifs ne constitue pas une solution intéressante pour certaines applications, telles que la détection d'intrusions, la détection des fraudes, l'audit des risques, l'identification des valeurs extrêmes dans les bases de données, l'analyse des données criminelles, l'analyse du désordre génétique à partir des données biologiques, l'analyse des maladies rares à partir des données médicales, l'analyse des données d'apprentissage en ligne, etc. (Booker, 2009; He et Xu, 2005; Koh et Rountree, 2010; Mahmood et al., 2010; Manning et al., 2008; Romero et al., 2010; Szathmary et al., 2010). En effet, dans de telles situations, un comportement fréquent peut être sans valeur ajoutée pour l'utilisateur final. Par contre, les événements peu fréquents sont les plus intéressants parce qu'ils indiquent qu'un évènement inattendu, une exception par exemple (Taniar et al., 2008), est survenue. Une étude doit alors continuer afin de déterminer les causes possibles de ce changement peu commun du comportement normal. La fouille des motifs rares s'est alors avérée d'une réelle valeur ajoutée (Koh

Adresse IP	Port	Authentification	Date
197.2.123.87	23	NV	d_1
197.1.104.19	1221	NV	d_2
194.23.22.2	80	V	d_3
197.1.104.19	225	V	d_4
197.2.123.29	21	V	d_5
197.1.104.19	1221	NV	d_6
⋮	⋮	⋮	⋮
197.1.156.27	145	V	d_n

TAB. 1 – *Extrait d'un fichier log d'accès à un serveur Web.*

et Rountree, 2010; Weiss, 2004). En effet, ces motifs, ayant une fréquence d'apparition dans la base inférieure à un certain seuil donné, permettent de cerner les évènements rares, peu communs, inattendus, exceptionnels, cachés, etc. (Berberidis et Vlahavas, 2007; Padmanabhan et Tuzhilin, 2006; Weiss, 2004). Comme illustration des applications des motifs rares dans le domaine de la sécurité informatique, étant donné un fichier log qui représente les tentatives de connexions effectuées sur un serveur Web d'authentification, ces motifs véhiculent les informations liées aux tentatives d'attaques à savoir par exemple l'origine des attaques, les ports les plus attaqués et les services les plus visés. Considérons par exemple la table 1 qui représente un échantillon réduit d'un tel fichier où V dénote Valide, NV dénote Non Valide, et d_i une date d'accès. Par exemple, si le motif (*197.1.104.19*, 1221, NV) s'avère rare, l'adresse *197.1.104.19* peut être considérée à l'origine d'une attaque sur le port 1221. Une analyse détaillée de ses accès est alors à effectuer.

Dans la pratique, l'exploitation des motifs rares est confrontée à diverses contraintes dont les principales sont : (i) l'extraction complexe de ces motifs qui ne bénéficient pas des propriétés des motifs fréquents et par conséquent les critères d'élagage appliqués pour ces derniers ne sont pas exploitables ; (ii) le nombre très important des motifs rares dans les applications réelles ; et (iii) la qualité des motifs rares extraits et qui peuvent comporter des items qui n'ont aucun lien sémantique entre eux. Par exemple, le motif composé par les items "$Lait$" et "$Caviar$" est un motif rare. Cependant, aucune corrélation n'existe entre le produit "$Lait$" très fréquemment acheté et le produit "$Caviar$" cher et rarement acheté.

Afin que l'exploitation des motifs extraits soit fructueuse, leur nombre relativement réduit et leur qualité intéressante sont deux critères importants que doit chercher à faire émerger un processus de fouille. Dans le domaine médical ou encore dans la sécurité des réseaux informatiques par exemple, une information exacte et précise est exigée. Ainsi, l'idée d'extraire les motifs rares tout en intégrant les mesures de corrélations est d'une grande utilité. En effet, l'intégration de telles mesures permet de limiter l'ensemble extrait aux motifs rares ayant une corrélation entre leurs items dépassant un certain seuil de corrélation. Ces motifs *corrélés rares* offrent ainsi un fort lien sémantique entre les items les composant.

Dans (Hamrouni et al., 2011), les auteurs ont proposé des représentations concises *sans perte d'information*, appelées aussi *exactes*, des motifs rares sans aucune considération des mesures de corrélations dans le processus de fouille. Une étude des représentations concises des motifs fréquents (Calders et al., 2005) a été alors menée afin de proposer celles des motifs

rares. Elle a prouvé l'intérêt de considérer la notion de classe d'équivalence, associée à l'opérateur de fermeture de Galois (Ganter et Wille, 1999), permettant de réduire la redondance au sein des motifs en regroupant ensemble ceux caractérisant un même ensemble de transactions. Les éléments minimaux et maximaux de la classe, les générateurs minimaux (appelé aussi itemsets libres) et les itemsets fermés (Pasquier et al., 2005) respectivement, sont ainsi à la base des représentations des motifs rares proposées. Par ailleurs, un des résultats clés de cette étude est que les représentations basées sur les règles de déduction (Calders et al., 2005) et celles basées sur les identités d'inclusion-exclusion (Casali et al., 2005; Hamrouni et al., 2009) ne sont pas adaptées à la fouille des motifs rares.

D'autre part, l'approche présentée dans (Ben Younes et al., 2010) utilise la mesure de corrélation *bond* (Omiecinski, 2003) pour l'extraction de représentations concises exactes des motifs corrélés fréquents. Ceci a permis de ne retenir qu'un sous-ensemble des motifs fréquents, constitué par les motifs présentant une forte corrélation entre les items les constituant. Le choix de cette mesure a été effectué sur la base d'une étude de ses propriétés qui se sont avérées plus intéressantes que celles d'autres mesures de corrélation. Toutefois, l'intégration d'une mesure de corrélation est d'une utilité encore plus grande dans le cas de la fouille des motifs rares. En effet, elle permet d'éviter l'extraction de motifs contenant des items n'ayant aucun lien sémantique entre eux ce qui expliquerait en quelque sorte pourquoi ces motifs sont rares. Ainsi, sans l'utilisation d'une mesure de corrélation, un motif rare peut ne représenter aucune information utile s'il est composé d'items faiblement corrélés entre eux. Un motif rare intéressant serait donc celui qui apparaît un nombre très faible de fois dans la base tout en ayant des items qui sont fortement liés, *c.-à.-d.* que l'apparition de l'un dépend de celles des autres.

Ainsi, dans cet article, nous allons nous intéresser à l'extraction des représentations concises exactes des motifs corrélés rares. Dans ce cadre, nous nous intéressons à la mesure de corrélation *bond* correspondant au rapport entre le support conjonctif d'un motif et son support disjonctif. Notre choix de cette mesure est motivé par le cadre théorique dont elle bénéficie (Omiecinski, 2003) ainsi que l'étude structurelle qui a été effectuée dans (Ben Younes et al., 2010). En plus, il a été prouvé dans (Surana et al., 2010) que la mesure *bond* vérifie les propriétés théoriques que toute mesure de qualité dédiée aux règles d'association rares doit avoir. Par ailleurs, dans (Segond et Borgelt, 2011), les auteurs ont proposé une approche générique de fouille des motifs corrélés. La mesure de corrélation *bond* a été utilisée ainsi que onze autres mesures de corrélations vérifiant toutes la propriété d'anti-monotonie. L'extraction des motifs corrélés a été alors montrée plus complexe tout en étant plus informative que celle des motifs fréquents (Segond et Borgelt, 2011). Il est toutefois important de noter qu'aucune étude de l'ensemble des motifs corrélés rares n'a été effectuée dans (Segond et Borgelt, 2011; Surana et al., 2010). Grâce à cette propriété d'anti-monotonie, les motifs corrélés selon la mesure *bond* induisent un idéal d'ordre dans le treillis des motifs (tout sous-ensemble d'un motif corrélé est aussi corrélé). Par opposition à ces derniers, les motifs rares induisent un filtre d'ordre et vérifient une contrainte monotone (tout sur-ensemble d'un motif rare est aussi rare). Par conséquent, l'ensemble des motifs corrélés rares que l'on vise à extraire résulte de l'intersection des deux théories (Mannila et Toivonen, 1997) associées respectivement aux contraintes de corrélation et de rareté.

La nature opposée de ces contraintes permet de différencier l'ensemble des motifs corrélés rares de l'ensemble des motifs induit par une ou plusieurs contraintes de même type (Bouli-

caut et Jeudy, 2010; Pei et Han, 2004). Cette caractéristique rend plus complexe l'extraction de l'ensemble des motifs corrélés rares. À cet égard, nous proposons dans cet article une caractérisation de cet ensemble moyennant la notion de classe d'équivalence. Dans notre cas, les classes d'équivalence seront induites par l'opérateur de fermeture associé à la mesure de corrélation *bond*. Ces classes jouent un rôle clé dans l'élimination de la redondance entre les motifs. Une fois la caractérisation effectuée, nous proposons des représentations exactes des motifs corrélés rares. Ces représentations permettent d'une part de réduire significativement le nombre de motifs corrélés rares extraits. Elles améliorent aussi leur qualité et ce en évitant la redondance entre les motifs puisqu'elles ne maintiennent qu'un sous-ensemble sans perte d'information de l'ensemble total des motifs corrélés rares. D'autre part, elles assurent la régénération aisée et efficace de l'ensemble des motifs corrélés rares. Il est aussi important de noter que les représentations proposées permettent non seulement la dérivation du support conjonctif mais aussi des supports disjonctif et négatif des motifs corrélés rares. Ceci permet par exemple d'utiliser ces motifs comme base pour l'extraction des règles généralisées où les connecteurs de disjonction et de négation sont utilisés en plus de celui classique de conjonction (Hamrouni et al., 2010).

Au meilleur de notre connaissance, aucune étude n'a été réalisée dans la littérature dans le but de proposer une représentation concise des motifs corrélés rares. Par ailleurs, une autre originalité de ce travail consiste en la nature des contraintes manipulées et la caractérisation de l'ensemble résultant moyennant une relation d'équivalence. Nous signalons aussi que l'approche proposée dans ce travail n'est pas restreinte aux motifs corrélés rares selon cette mesure. En effet, elle est générique dans le sens qu'elle s'applique à tout ensemble de motifs corrélés rares selon toute mesure de corrélation vérifiant les mêmes propriétés structurelles que la mesure *bond* telle que par exemple la mesure *all-confidence* (Omiecinski, 2003) [1]. Il est toutefois à noter que la mesure *bond* a pour avantage de permettre la dérivation des supports disjonctif et négatif, ce que ne peuvent pas permettre les autres mesures. Notons que dans cet article, nous nous focalisons principalement sur la caractérisation des représentations proposées. L'étude détaillée de l'aspect algorithmique associé à l'extraction des représentations proposées fera l'objet d'un futur travail.

Le reste de l'article est organisé comme suit : dans la section 2, nous présentons les notions de base qui seront utilisées tout au long de ce travail. Dans la section 3, nous caractérisons l'ensemble des motifs corrélés rares en étudiant les contraintes associées. Nous discutons aussi dans cette section différentes approches de l'état de l'art traitant de la conjonction de contraintes monotones et de contraintes anti-monotones. La section 4 présente une description détaillée des représentations concises proposées des motifs corrélés rares. Les conclusions et les perspectives de travaux futurs seront présentées dans la section 5.

2 Concepts de base

2.1 Extraction de motifs

Nous commençons par présenter l'ensemble des notions de base relatives à l'extraction des motifs. Définissons tout d'abord une base de transactions.

[1] Mathématiquement équivalente à la mesure *h-confidence* (Xiong et al., 2006).

Définition 1 (***Base de transactions***) *Une base de transactions* (*appelée aussi contexte d'extraction ou simplement contexte*) *est représentée sous la forme d'un triplet* $\mathcal{D} = (\mathcal{T}, \mathcal{I}, \mathcal{R})$ *dans lequel* $\mathcal{T}$ *et* $\mathcal{I}$ *sont, respectivement, des ensembles finis de transactions* (*ou objets*) *et d'items* (*ou attributs*), *et* $\mathcal{R} \subseteq \mathcal{T} \times \mathcal{I}$ *est une relation binaire entre les transactions et les items. Un couple* $(t, i) \in \mathcal{R}$ *dénote le fait que la transaction* $t \in \mathcal{T}$ *contient l'item* $i \in \mathcal{I}$.

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

TAB. 2 – *Un exemple d'une base de transactions.*

Dans ce travail, nous nous sommes principalement intéressés aux itemsets comme classe de motifs. Les supports d'un motif sont définis comme suit :

Définition 2 (***Supports d'un motif***) *Soient* $\mathcal{D} = (\mathcal{T}, \mathcal{I}, \mathcal{R})$ *une base de transactions et un motif non vide* $I \subseteq \mathcal{I}$. *Nous distinguons trois types de supports correspondants à* I *:*

- ***Le support conjonctif :*** *Supp*$(\wedge I) = |\{t \in \mathcal{T} \mid \forall\, i \in I : (t, i) \in \mathcal{R}\}|$
- ***Le support disjonctif :*** *Supp*$(\vee I) = |\{t \in \mathcal{T} \mid \exists\, i \in I : (t, i) \in \mathcal{R}\}|$
- ***Le support négatif :*** *Supp*$(\neg I) = |\{t \in \mathcal{T} \mid \forall\, i \in I : (t, i) \notin \mathcal{R}\}|$

Exemple 1 *Considérons la base de transactions illustrée par la table 2 et qui sera utilisée dans la suite pour les différents exemples. Nous avons Supp*$(\wedge AD) = |\{1\}| = 1$, *Supp*$(\vee AD) = |\{1, 3, 5\}| = 3$, *et, Supp*$(\neg(AD)) = |\{2, 4\}| = 2$ [(2)].

Dans la suite, s'il n'y a pas de risque de confusion, le *support conjonctif* sera simplement appelé *support*.

La définition suivante présente le statut de fréquence d'un motif, fréquent ou infréquent, étant donné un seuil minimal de support (Agrawal et Srikant, 1994).

Définition 3 (***Motif fréquent/rare***) *Soit une base de transactions* $\mathcal{D} = (\mathcal{T}, \mathcal{I}, \mathcal{R})$, *un seuil minimal de support conjonctif minsupp, un motif* $I \subseteq \mathcal{I}$ *est dit fréquent si Supp*$(\wedge I) \geq$ *minsupp.* I *est dit infréquent ou rare sinon.*

Exemple 2 *Soit minsupp = 2. Supp*$(\wedge BCE) = 3$, *le motif* BCE *est un motif fréquent. Cependant, le motif* CD *est non fréquent ou rare puisque Supp*$(\wedge CD) = 1 < 2$.

Outre la contrainte de fréquence minimale traduite par le seuil *minsupp*, d'autres contraintes peuvent être intégrées dans le processus d'extraction des motifs. Ces contraintes admettent différents types, dont les deux principaux sont définis dans ce qui suit (Boulicaut et Jeudy, 2010; Pei et Han, 2004).

[2]Nous employons une forme sans séparateur pour les ensembles d'items : par exemple, AD représente l'ensemble {A, D}.

Définition 4 (***Contrainte anti-monotone/monotone***)
- *Une contrainte Q est anti-monotone si* $\forall\ I \subseteq \mathcal{I}, \forall\ I_1 \subseteq I$ *: I satisfait $Q \Rightarrow I_1$ satisfait Q.*
- *Une contrainte Q est monotone si* $\forall\ I \subseteq \mathcal{I}, \forall\ I_1 \supseteq I$ *: I satisfait $Q \Rightarrow I_1$ satisfait Q.*

Exemple 3 *La contrainte de fréquence, c.-à.-d. avoir un support supérieur ou égal à minsupp, est une contrainte anti-monotone. En effet,* $\forall\ I, I_1 \subseteq \mathcal{I}$*, si* $I_1 \subseteq I$ *et Supp*$(\wedge I) \geq$ *minsupp, alors Supp*$(\wedge I_1) \geq$ *minsupp puisque Supp*$(\wedge I_1) \geq$ *Supp*$(\wedge I)$.

D'une manière duale, la contrainte de rareté, c.-à.-d. avoir un support strictement inférieur à minsupp, est monotone. En effet, $\forall\ I, I_1 \subseteq \mathcal{I}$*, si* $I_1 \supseteq I$ *et Supp*$(\wedge I) <$ *minsupp, alors Supp*$(\wedge I_1) <$ *minsupp puisque Supp*$(\wedge I_1) \leq$ *Supp*$(\wedge I)$.

Soit $\mathcal{P}(\mathcal{I})$ l'ensemble de tous les sous-ensembles de $\mathcal{I}$. Dans ce qui suit, nous introduisons les notions duales d'idéal d'ordre et de filtre d'ordre (Ganter et Wille, 1999) définis sur $\mathcal{P}(\mathcal{I})$.

Définition 5 (***Idéal d'ordre***) *Un sous-ensemble $\mathcal{S}$ de $\mathcal{P}(\mathcal{I})$ est un idéal d'ordre s'il vérifie les propriétés suivantes :*
- *Si* $I \in \mathcal{S}$*, alors* $\forall\ I_1 \subseteq I : I_1 \in \mathcal{S}$.
- *Si* $I \notin \mathcal{S}$*, alors* $\forall\ I \subseteq I_1 : I_1 \notin \mathcal{S}$.

Définition 6 (***Filtre d'ordre***) *Un sous-ensemble $\mathcal{S}$ de $\mathcal{P}(\mathcal{I})$ est un filtre d'ordre s'il vérifie les propriétés suivantes :*
- *Si* $I \in \mathcal{S}$*, alors* $\forall\ I_1 \supseteq I : I_1 \in \mathcal{S}$.
- *Si* $I \notin \mathcal{S}$*, alors* $\forall\ I \supseteq I_1 : I_1 \notin \mathcal{S}$.

Une contrainte anti-monotone telle que la contrainte de fréquence induit un idéal d'ordre. D'une manière duale, une contrainte monotone telle que la contrainte de rareté forme un filtre d'ordre.

L'ensemble des motifs satisfaisant une contrainte donnée est appelé *théorie* (Mannila et Toivonen, 1997). Comme dans ce travail, nous nous intéressons principalement aux motifs rares qui sont aussi corrélés, nous présentons certaines propriétés utiles des motifs rares et des motifs corrélés. En particulier, nous nous intéressons aux bordures délimitant respectivement ces deux ensembles de motifs.

2.2 Motifs rares

L'ensemble des motifs rares, correspondant aux motifs vérifiant la contrainte de rareté, est défini comme suit :

Définition 7 (***Motifs rares***) *L'ensemble des motifs rares est défini par :* $\mathcal{MR} = \{I \subseteq \mathcal{I} |$ *Supp*$(\wedge I) <$ *minsupp*$\}$.

Comme indiqué plus haut, cet ensemble forme un filtre d'ordre dans $\mathcal{P}(\mathcal{I})$. Il en résulte que tous les sur-ensembles d'un motif rare sont aussi rares.

Exemple 4 *Soit minsupp = 4. Le motif* `BC` *est rare puisque Supp*$(\wedge$`BC`$) = 3 < 4$*. Par conséquent,* `BC` $\in \mathcal{MR}$*. Ainsi,* `ABC` $\in \mathcal{MR}$ *puisque* `BC` $\subseteq$ `ABC`*. En effet, Supp*$(\wedge$`ABC`$) = 2 < 4$.

Dans la suite, nous aurons besoin des motifs rares les plus petits, par rapport à la relation d'inclusion ensembliste. Ces motifs forment la bordure positive des motifs rares, *c.-à.-d.* ceux rares et dont tous les sous-ensembles sont fréquents et sont définis comme suit :

Définition 8 (***Motifs rares minimaux***) *L'ensemble $\mathcal{MRM}in$ des motifs rares minimaux correspond aux motifs rares n'ayant aucun sous-ensemble strict rare. Cet ensemble est défini par : $\mathcal{MRM}in$ = $\{I \in \mathcal{MR} | \forall\ I_1 \subset I : I_1 \notin \mathcal{MR}\}$, ou d'une manière équivalente : $\mathcal{MRM}in$ = $\{I \in \mathcal{MR} | \forall\ I_1 \subset I$: Supp($\wedge I_1$) $\geq$ minsupp$\}$.*

Exemple 5 *Considérons la base illustrée par la table 2 pour minsupp = 4. Le motif* `BC` $\in \mathcal{MRM}in$ *puisque Supp($\wedge$*`BC`*) = 3 < 4 et, d'autre part, Supp($\wedge$*`B`*) = Supp($\wedge$*`C`*) = 4. Il en est de même pour* `A`*,* `D` *et* `CE`*. Ainsi, dans ce cas, $\mathcal{MRM}in$ = $\{$*`A`*,* `D`*,* `BC`*,* `CE`$\}$.

Après avoir présenté les motifs rares, nous allons étudier dans ce qui suit les propriétés des motifs corrélés selon la mesure *bond*.

2.3 Motifs corrélés selon la mesure *bond*

La mesure *bond* (Omiecinski, 2003) est mathématiquement équivalente aux mesures *coherence* (Lee et al., 2003), *coefficient de Tanimoto* (Tanimoto, 1958) et *Jaccard* (Jaccard, 1901). Elle a été redéfinie dans (Ben Younes et al., 2010) comme suit :

Définition 9 (***Mesure bond***) *Soit $I \subseteq \mathcal{I}$. La mesure bond de I est définie par :*

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

La mesure *bond* prend ses valeurs sur l'intervalle [0, 1]. En considérant *l'univers* d'un motif I (Lee et al., 2003), *c.-à.-d.* l'ensemble des transactions contenant un sous-ensemble non vide de I, la mesure *bond* véhicule l'information concernant le taux d'apparition simultanée des items d'un motif dans son univers. Ainsi, plus les items du motif sont dispersés dans son univers (*c.-à.-d.* faiblement corrélés), plus sera faible la valeur de *bond* puisque *Supp*($\wedge I$) serait nettement plus petit que *Supp*($\vee I$). Inversement, plus les items de I dépendent les uns des autres (*c.-à.-d.* fortement corrélés), plus sera élevée la valeur de *bond* puisque *Supp*($\wedge I$) serait proche de *Supp*($\vee I$).

Nous définissons maintenant les motifs corrélés selon cette mesure.

Définition 10 (***Motifs corrélés selon la mesure bond***) *Soit minbond un seuil minimal de corrélation. L'ensemble $\mathcal{MC}$ des motifs corrélés selon la mesure bond est défini par : $\mathcal{MC}$ = $\{I \subseteq \mathcal{I} \mid bond(I) \geq minbond\}$*

Exemple 6 *Considérons la base illustrée par la table 2 pour minbond = 0,5. Nous avons bond(*`AB`*) = $\frac{2}{5}$ = 0,4 < 0,5. Le motif* `AB` *est alors non corrélé. Par contre, bond(*`BCE`*) = $\frac{3}{5}$ = 0,6 $\geq$ 0,5. Ainsi, le motif* `BCE` *est corrélé.*

Il a été démontré dans (Ben Younes et al., 2010) que la mesure *bond* présente diverses propriétés intéressantes. En effet, cette mesure : (i) est *symétrique* puisque $\forall\ I, J \subseteq \mathcal{I}$, *bond*($IJ$) = *bond*($JI$) ; ($ii$) est *descriptive c.-à.-d.* insensible au changement du nombre de transactions ; (iii) vérifie la propriété de *cross support* (Xiong et al., 2006). Grâce à cette dernière propriété, pour un motif $I \subseteq \mathcal{I}$ et pour un seuil minimal de corrélation *minbond*, s'il existe un couple d'items $x, y \in I$ tel que $\frac{Supp(\wedge x)}{Supp(\wedge y)} < minbond$, alors I n'est pas corrélé puisque *bond*(I) <

minbond. I vérifie alors la propriété de cross-support pour le seuil *minbond* ; et (iv) induit une contrainte anti-monotone du moment que le seuil minimal *minbond* est fixé. En effet, $\forall\ I,\ I_1 \subseteq \mathcal{I}$, si $I_1 \subseteq I$, alors *bond*(I_1) $\geq$ *bond*(I). Ainsi, l'ensemble $\mathcal{MC}$ des motifs corrélés forme un idéal d'ordre. Autrement dit, si un motif est corrélé, alors tous ses sous-ensembles sont aussi corrélés.

La proposition suivante présente une relation intéressante entre la valeur de la mesure *bond* et les valeurs des supports conjonctifs et disjonctifs pour chaque couple de deux motifs I et I_1 tel que $I \subseteq I_1$ (Ben Younes et al., 2010).

Proposition 1 *Soient $I, I_1 \subseteq \mathcal{I}$ et $I \subseteq I_1$. Si bond(I) = bond(I_1), alors Supp($\wedge I$) = Supp($\wedge I_1$) et Supp($\vee I$) = Supp($\vee I_1$).*

D'après la proposition précédente, si *bond*(I) = *bond*(I_1), alors *Supp*($\neg I$) = *Supp*($\neg I_1$). En effet, I et I_1 ont le même support disjonctif et, par la loi de De Morgan, nous avons le lien suivant entre les supports disjonctif et négatif d'un motif : *Supp*($\neg I$) = $|\mathcal{T}|$ - *Supp*($\vee I$). D'autre part, si *bond*(I) $\neq$ *bond*(I_1), alors *Supp*($\wedge I$) $\neq$ *Supp*($\wedge I_1$) ou *Supp*($\vee I$) $\neq$ *Supp*($\vee I_1$) (*c.-à.-d.* un des deux supports est différent ou les deux à la fois).

Dans la suite, nous aurons besoin de l'ensemble des motifs corrélés maximaux défini formellement comme suit.

Définition 11 (***Motifs corrélés maximaux***) *L'ensemble des motifs corrélés maximaux constitue la bordure positive des motifs corrélés et correspond aux motifs corrélés n'admettant aucun sur-ensemble strict corrélé. Cet ensemble est défini par : $\mathcal{MCM}ax$ = $\{I \in \mathcal{MC} | \forall\ I_1 \supset I$: $I_1 \notin \mathcal{MC}\}$, ou d'une manière équivalente : $\mathcal{MCM}ax$ = $\{I \in \mathcal{MC} | \forall\ I_1 \supset I$: bond(I_1) $<$ minbond$\}$.*

Exemple 7 *Soit la base illustrée par la table 2. Pour minbond = 0,2, nous avons $\mathcal{MCM}ax$ = $\{$`ACD`, `ABCE`$\}$. En effet, quelque soit le sur-ensemble strict de `ACD` ou de `ABCE`, ce sur-ensemble n'est pas corrélé.*

La sous-section suivante est dédiée à la présentation de l'opérateur de fermeture associé à la mesure *bond*. Cet opérateur permettra de caractériser les motifs grâce aux classes d'équivalence induites.

2.4 Opérateur de fermeture f_{bond} associé à la mesure *bond*

L'opérateur de fermeture associé à la mesure *bond* est défini comme suit (Ben Younes et al., 2010) :

Définition 12 (***Opérateur*** f_{bond}) *L'opérateur associé à la mesure bond est défini comme suit :*

$$\begin{array}{rcl} f_{bond} : \mathcal{P}(\mathcal{I}) & \rightarrow & \mathcal{P}(\mathcal{I}) \\ I & \mapsto & f_{bond}(I) = I \cup \{i \in \mathcal{I} \setminus I |\ bond(I) = bond(I \cup \{i\})\} \end{array}$$

L'opérateur f_{bond} a été démontré d'être un opérateur de fermeture (Ben Younes et al., 2010). En effet, il vérifie les propriétés d'extensivité, d'isotonie, et d'idempotence (Ganter et Wille, 1999). Le motif fermé d'un motif I par f_{bond}, *c.-à.-d.* $f_{bond}(I)$, est ainsi l'ensemble maximal d'items contenant I et ayant la même valeur de la mesure *bond* que I.

Exemple 8 *Soit la base illustrée par la table 2. Pour minbond = 0,2, nous avons bond(*$\mathtt{AB}$*) =* $\frac{2}{5}$*, bond(*$\mathtt{ABC}$*) =* $\frac{2}{5}$*, bond(*$\mathtt{ABE}$*) =* $\frac{2}{5}$*. Ainsi,* $\mathtt{C} \in f_{bond}(\mathtt{AB})$*, et* $\mathtt{E} \in f_{bond}(\mathtt{AB})$*. Par contre, bond(*$\mathtt{ABD}$*) =* $\frac{0}{5}$ *= 0. Ainsi,* $\mathtt{D} \notin f_{bond}(\mathtt{AB})$*. Par conséquent,* $f_{bond}(\mathtt{AB}) = \mathtt{ABCE}$*.*

L'application de l'opérateur f_{bond} partitionne l'ensemble des parties de $\mathcal{I}$ en des classes d'équivalence disjointes définies comme suit.

Définition 13 (***Classe d'équivalence associée à l'opérateur de fermeture*** f_{bond}) *Une classe d'équivalence associée à l'opérateur de fermeture* f_{bond} *contient un ensemble de tous les motifs possédant la même fermeture par* f_{bond}*.*

Chaque classe d'équivalence est caractérisée par un élément maximal – un motif fermé – et un ou plusieurs éléments minimaux – des motifs minimaux corrélés. Nous définissons formellement ces motifs.

Définition 14 (***Motifs fermés corrélés***) *L'ensemble* $\mathcal{MFC}$ *des motifs fermés corrélés par* f_{bond} *est défini par :* $\mathcal{MFC} = \{I \in \mathcal{MC} | \nexists\, I_1 \supset I : bond(I) = bond(I_1)\}$*, ou d'une manière équivalente :* $\mathcal{MFC} = \{I \in \mathcal{MC} | \nexists\, I_1 \supset I : f_{bond}(I) = f_{bond}(I_1)\}$*.*

Exemple 9 *Soit la base illustrée par la table 2 pour minbond = 0,2. Le motif* $\mathtt{ACD}$ *est corrélé puisque bond(*$\mathtt{ACD}$*) =* $\frac{1}{4}$ *= 0,25* $\geq$ *0,2. Il est aussi fermé puisqu'il n'admet pas de sur-ensemble strict de même valeur de bond. En effet, bond(*$\mathtt{ABCD}$*) = 0, bond(*$\mathtt{ACDE}$*) = 0, et par conséquent bond(*$\mathtt{ABCDE}$*) = 0.*

Définition 15 (***Motifs minimaux corrélés***) *L'ensemble* $\mathcal{MMC}$ *des motifs minimaux corrélés est défini par :* $\mathcal{MMC} = \{I \in \mathcal{MC} | \nexists\, I_1 \subset I : bond(I) = bond(I_1)\}$*, ou d'une manière équivalente :* $\mathcal{MMC} = \{I \in \mathcal{MC} | \nexists\, I_1 \subset I : f_{bond}(I) = f_{bond}(I_1)\}$*.*

Exemple 10 *Soit la base illustrée par la table 2 pour minbond = 0,2. Le motif* $\mathtt{AB}$ *est corrélé puisque bond(*$\mathtt{AB}$*) =* $\frac{2}{5}$ *= 0,4 > 0,2. Il est aussi minimal puisque bond(*$\mathtt{A}$*) = bond(*$\mathtt{B}$*) = 1.*

L'ensemble $\mathcal{MMC}$ des motifs minimaux corrélés forme un idéal d'ordre. En effet, cet ensemble contient les motifs vérifiant la contrainte anti-monotone "être minimal dans sa classe d'équivalence et être corrélé". En effet, cette dernière résulte de la conjonction de deux contraintes anti-monotones, à savoir "être un motif minimal" et "être un motif corrélé".

La proposition suivante indique les propriétés communes à deux motifs appartenant à une même classe d'équivalence induite par f_{bond}.

Proposition 2 *Soit* $\mathcal{C}$ *une classe d'équivalence associée à l'opérateur de fermeture* f_{bond} *et* I *et* $I_1 \in \mathcal{C}$*. Nous avons :* ***a)*** $f_{bond}(I) = f_{bond}(I_1)$*,* ***b)*** *bond*$(I)$ = *bond*(I_1)*,* ***c)*** *Supp*$(\wedge I)$ = *Supp*$(\wedge I_1)$*,* ***d)*** *Supp*$(\vee I)$ = *Supp*$(\vee I_1)$*, et,* ***e)*** *Supp*$(\neg I)$ = *Supp*$(\neg I_1)$*.*

Preuve.

a) Grâce à la définition 13, I et I_1 ont la même fermeture par f_{bond}. Soit F cette fermeture.

b) Comme l'opérateur de fermeture préserve la valeur de la mesure *bond* d'un motif (*cf.* Définition 12), et puisque I et I_1 ont la même fermeture F, nous avons *bond*(I) = *bond*(F), et *bond*(I_1) = *bond*(F). Ainsi, *bond*(I) = *bond*(I_1).

c), d), et e) Comme $I \subseteq F$ et *bond*(I) = *bond*(F), d'après la proposition 1, I et F admettent les mêmes supports conjonctif, disjonctif, et négatif. Il en est de même pour I_1 et F. Ainsi, I et I_1 ont les mêmes supports conjonctif, disjonctif, et négatif. $\diamondsuit$

Exemple 11 *Soit la base illustrée par la table 2 et minbond = 0,2. Considérons la classe d'équivalence dont le motif fermé corrélé est* ABCE*. Les motifs minimaux corrélés associés sont* AB *et* AE*. Les motifs corrélés, non fermés ni minimaux, sont* ABE*,* ABC*, et* ACE*. Chacun de ces derniers est compris entre un motif minimal et le fermé corrélé. Tous les motifs de cette classe d'équivalence partagent la même valeur de la mesure bond égale à* $\frac{2}{5}$*, le même support conjonctif égal à 2, le même support disjonctif égal à 5, et le même support négatif égal à 0.*

Ainsi, tous les motifs d'une classe d'équivalence induite par f_{bond} apparaissent dans les mêmes transactions (grâce à l'égalité du support conjonctif). En plus, les items associés aux motifs de la classe caractérisent les mêmes transactions. En effet, chacune de ces dernières contient nécessairement un sous-ensemble non vide de chaque motif de la classe (grâce à l'égalité du support disjonctif). Cet opérateur de fermeture lie ainsi l'espace de recherche conjonctif et celui disjonctif (Ben Younes et al., 2010). Le motif fermé de la classe offre ainsi l'expression la plus spécifique caractérisant ces transactions, tandis qu'un des motifs minimaux représente une des expression les plus générales.

Dans la section suivante, nous allons nous focaliser sur les motifs corrélés rares.

3 Caractérisation de l'ensemble des motifs corrélés rares

3.1 Définition et propriétés

L'ensemble des motifs corrélés rares est défini comme suit.

Définition 16 (***Motifs corrélés rares associés à la mesure bond***) *Étant donnés les seuils minimaux de support conjonctif et de corrélation minsupp et minbond, respectivement, l'ensemble des motifs corrélés rares, dénoté* $\mathcal{MCR}$*, est défini comme suit :* $\mathcal{MCR} = \{I \subseteq \mathcal{I} |$ *Supp*$(\wedge I) <$ *minsupp et bond*$(I) \geq$ *minbond*$\}$.

Exemple 12 *Considérons la base illustrée par la table 2 pour minsupp = 4 et minbond = 0,2. L'ensemble* $\mathcal{MCR}$ *est composé des motifs suivants où chaque triplet représente le motif, sa valeur de support et sa valeur de bond :* $\mathcal{MCR} = \{$(A, 3, $\frac{3}{3}$), (D, 1, $\frac{1}{1}$), (AB, 2, $\frac{2}{5}$), (AC, 3, $\frac{3}{4}$), (AD, 1, $\frac{1}{3}$), (AE, 2, $\frac{2}{5}$), (BC, 3, $\frac{3}{5}$), (CD, 1, $\frac{1}{4}$), (CE, 3, $\frac{3}{5}$), (ABC, 2, $\frac{2}{5}$), (ABE, 2, $\frac{2}{5}$), (ACD, 1, $\frac{1}{4}$), (ACE, 2, $\frac{2}{5}$), (BCE, 3, $\frac{3}{5}$), (ABCE, 2, $\frac{2}{5}$)$\}$. *Cet ensemble est schématisé par la figure 1. Le support indiqué en haut à gauche de chaque cadre représentant un motif est son support conjonctif. Comme le montre cette figure, l'ensemble* $\mathcal{MCR}$ *des motifs corrélés rares correspond aux motifs localisés en dessous de la bordure de la contrainte anti-monotone*

associée aux motifs corrélés, et au dessus de la bordure de la contrainte monotone associée aux motifs rares.

Il résulte de la définition précédente que $\mathcal{MCR}$ correspond à l'intersection de l'ensemble des motifs corrélés et de l'ensemble des motifs rares : $\mathcal{MCR} = \mathcal{MC} \cap \mathcal{MR}$. La proposition suivante découle de ce résultat.

Proposition 3 *Soit $I \in \mathcal{MCR}$. Nous avons :*
- *D'après l'idéal d'ordre de l'ensemble des motifs corrélés selon la mesure bond, $\forall\ I_1 \subseteq I : I_1 \in \mathcal{MC}$*
- *D'après le filtre d'ordre de l'ensemble des motifs rares, $\forall\ I_1 \supseteq I : I_1 \in \mathcal{MR}$.*

La preuve découle des propriétés induites par les contraintes de corrélation et de rareté. L'ensemble $\mathcal{MCR}$, dont les éléments vérifient la contrainte "être un motif corrélé rare", résulte ainsi de l'intersection de deux ordres résultant de deux contraintes de natures opposées. Cet ensemble n'est ainsi ni un idéal ni un filtre d'ordre. Dans le treillis des motifs, l'espace de recherche des motifs corrélés rares est ainsi délimité, d'une part, par les éléments maximaux vérifiant la contrainte de corrélation et qui sont rares, *c.-à.-d.* les motifs *rares* parmi *l'ensemble $\mathcal{MCM}ax$ des motifs corrélés maximaux* (*cf.* Définition 11) et, d'autre part, par les éléments minimaux vérifiant la contrainte de rareté et qui sont corrélés, *c.-à.-d.* les motifs corrélés parmi *l'ensemble $\mathcal{MRM}in$ des motifs rares minimaux* (*cf.* Définition 8). Ainsi, tout motif corrélé rare est nécessairement compris entre un élément de chacun des deux ensembles susmentionnés.

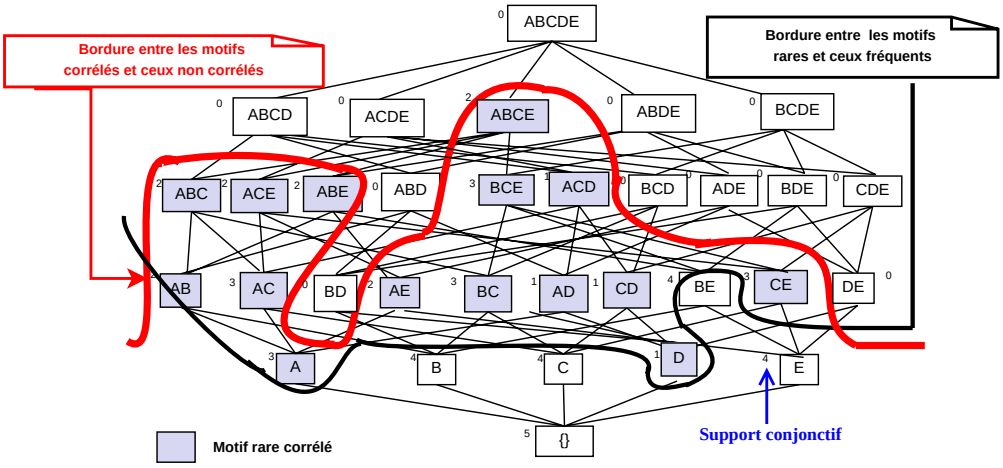

FIG. 1 – *Espace des motifs corrélés rares pour minsupp = 4 et minbond = 0,2.*

Exemple 13 *Considérons la figure 1 pour minsupp = 4 et minbond = 0,2. L'espace des motifs corrélés rares est délimité par : d'une part les motifs corrélés maximaux pour minbond = 0,2, à savoir* `ACD` *et* `ABCE` (*cf. Exemple 7*), *et, d'autre part, par les motifs rares minimaux pour minsupp = 4, à savoir* `A`, `D`, `BC` *et* `CE` (*cf. Exemple 5*). *Par exemple, le motif* `AD` *est un motif corrélé rare étant donné qu'il est compris entre un motif rare minimal à savoir* `D` *et un motif corrélé maximal à savoir* `ACD`.

Cet espace est ainsi plus difficile à localiser que les ensembles associés à des contraintes de même nature. En effet, la conjonction de contraintes anti-monotones (*resp.* monotones) est une contrainte anti-monotone (*resp.* monotone) (Bonchi et Lucchese, 2006). Par exemple, la contrainte "être un motif corrélé non rare (*c.-à.-d.* fréquent)" est une contrainte anti-monotone puisque résultante de la conjonction des contraintes anti-monotones "être un motif corrélé" et "être un motif fréquent". Elle induit donc un idéal d'ordre (Ben Younes et al., 2010). La contrainte "être un motif non corrélé rare" est une contrainte monotone et l'ensemble associé forme un filtre d'ordre dans le treillis des motifs.

D'un point de vue taille, et étant données les natures des contraintes induites par les seuils minimaux de support et de corrélation, à savoir respectivement *minsupp* et *minbond*, la taille de l'ensemble $\mathcal{MCR}$ des motifs corrélés rares varie de la manière indiquée dans la proposition suivante.

Proposition 4

a) *Soient minsupp*$_1$ *et minsupp*$_2$ *deux seuils minimaux de support et* $\mathcal{MCR}_{s1}$ *et* $\mathcal{MCR}_{s2}$ *les deux ensembles des motifs corrélés rares associés pour une même valeur de minbond. Nous avons : si minsupp*$_1 \leq$ *minsupp*$_2$*, alors* $\mathcal{MCR}_{s1} \subseteq \mathcal{MCR}_{s2}$ *et par conséquent* $|\mathcal{MCR}_{s1}| \leq |\mathcal{MCR}_{s2}|$.

b) *Soient minbond*$_1$ *et minbond*$_2$ *deux seuils minimaux de corrélation et* $\mathcal{MCR}_{b1}$ *et* $\mathcal{MCR}_{b2}$ *les deux ensembles des motifs corrélés rares associés pour une même valeur de minsupp. Nous avons : si minbond*$_1 \leq$ *minbond*$_2$*, alors* $\mathcal{MCR}_{b2} \subseteq \mathcal{MCR}_{b1}$ *et par conséquent* $|\mathcal{MCR}_{b2}| \leq |\mathcal{MCR}_{b1}|$.

Preuve.

- La preuve de **a)** dérive du fait que pour $I \subseteq \mathcal{I}$, si *Supp*$(\wedge I) <$ *minsupp* $_1$, alors *Supp*$(\wedge I) <$ *minsupp*$_2$. Ainsi, $\forall\ I \in \mathcal{MCR}_{s1}$, $I \in \mathcal{MCR}_{s2}$. Il en résulte que $\mathcal{MCR}_{s1} \subseteq \mathcal{MCR}_{s2}$.
- La preuve de **b)** dérive du fait que pour $I \subseteq \mathcal{I}$, si *bond*$(I) \geq$ *minbond* $_2$, alors *bond*$(I) \geq$ *minbond*$_1$. Ainsi, $\forall\ I \in \mathcal{MCR}_{b2}$, $I \in \mathcal{MCR}_{b1}$. Par conséquent, $\mathcal{MCR}_{b2} \subseteq \mathcal{MCR}_{b1}$. $\diamondsuit$

Ainsi, la taille de $\mathcal{MCR}$ est proportionnelle à *minsupp* et inversement proportionnelle à *minbond*. Il est toutefois à noter que, dans le cas général, nous ne pouvons rien décider quand les deux seuils varient en même temps et non seulement un à la fois.

Dans la suite, nous discutons les approches ayant été dédiées à l'extraction des motifs sous la conjonction de contraintes monotones et de contraintes anti-monotones.

3.2 Revue de l'état de l'art des approches traitant de la conjonction de contraintes monotones et de contraintes anti-monotones

Comme susmentionné, lors d'un processus de fouille de motifs, il est plus complexe de localiser les motifs vérifiant des contraintes de natures différentes que de localiser ceux associés à des contraintes de même nature. En effet, la nature variée des contraintes fait que les stratégies d'élagage ne sont applicables que pour une partie des contraintes et pas pour les autres. Ceci augmente le coût des traitements à effectuer afin de déterminer si un motif donné fait partie de l'ensemble recherché ou non.

Beaucoup de travaux ont été proposés dans la littérature afin de prendre en considération des contraintes de natures différentes lors du processus de fouille de motifs intéressants (Boulicaut et Jeudy, 2010; Pei et Han, 2004). Un des premiers algorithmes se situant dans ce cadre est

intitulé DUALMINER (Bucila et al., 2003). Cet algorithme effectue la réduction de l'espace de recherche en considérant à la fois les contraintes monotones et les contraintes anti-monotones. Cependant, DUALMINER souffre d'une limite liée au nombre élevé d'évaluation de contraintes (Boley et Gärtner, 2009). Dans (Lee et al., 2006), les auteurs proposent aussi une approche d'extraction des motifs fréquents sous contraintes. L'algorithme EXAMINER (Bonchi et al., 2005) a été aussi proposé pour la fouille des motifs fréquents sous une contrainte monotone. Il est toutefois important de noter que la stratégie de réduction du contexte d'extraction adoptée dans EXAMINER, bien qu'elle soit efficace, n'est pas applicable pour le cas d'une contrainte monotone sensible au changement du nombre de transactions. Ceci est le cas de la contrainte de rareté que nous traitons dans ce travail.

D'autres travaux ont été aussi dédiés à cette problématique, comme par exemple l'algorithme VST (De Raedt et al., 2002) d'extraction de chaînes de caractères sous la conjonction de contraintes de types différents. Cet algorithme parcourt l'espace de recherche par niveau à la APRIORI (Agrawal et Srikant, 1994). Il se déroule en deux phases principales : une phase de parcours du haut vers le bas pour l'élagage des motifs candidats selon la contrainte monotone et une phase de parcours du bas vers le haut permettant l'élagage des motifs candidats selon la contrainte anti-monotone. Par ailleurs, l'algorithme FAVST (Lee et De Raedt, 2004) a été introduit dans le but d'améliorer les performances de l'algorithme VST, en réduisant le nombre de balayages de la base. D'autres approches se situant dans ce cadre ont aussi vu le jour, tel que les algorithmes DPC-COFI (El-Hajj et Zaïane, 2005) et BIFOLDLEAP (El-Hajj et al., 2005). Ces approches mettent en place une stratégie qui consiste à extraire les motifs fréquents maximaux satisfaisant l'ensemble de contraintes posées puis à générer les sous-ensembles valides de ces motifs maximaux.

Diverses approches d'extraction des motifs corrélés sous contraintes ont été aussi proposées. Toutefois, la récupération de l'ensemble des motifs qui sont à la fois fortement corrélés et très peu fréquents est basée sur l'idée naïve d'extraire l'ensemble de tous les motifs fréquents pour un seuil *minsupp* très bas et, ensuite, de filtrer ces motifs par la contrainte de corrélation. Cette opération est très coûteuse en temps de traitement et en consommation de la mémoire à cause de l'explosion du nombre de candidats à évaluer. L'approche proposée dans (Sandler et Thomo, 2010) est fondée sur ce principe. Une autre stratégie d'extraction des motifs rares fortement corrélés, consiste à extraire l'ensemble de tous les motifs corrélés sans aucune intégration de la contrainte de support. L'ensemble récupéré englobe ainsi tous les motifs corrélés rares. L'approche proposée dans (Cohen et al., 2000) et l'algorithme DISCOVERMPATTERNS (Ma et Hellerstein, 2001) reposent sur ce principe. Il est ainsi à déduire que pour toutes ces approches, la contrainte monotone de rareté n'a été jamais incorporée dans la fouille afin de récupérer l'ensemble des motifs rares fortement corrélés.

L'approche proposée dans (Okubo et al., 2010) se situe aussi dans ce cadre. En effet, cette dernière est basée sur le principe qu'un motif faiblement corrélé par rapport à la mesure de corrélation *bond* est généralement rare dans le contexte d'extraction. La contrainte posée correspond à une restriction de la valeur de corrélation maximale. Cette dernière est monotone puisqu'elle correspond à l'opposée de la contrainte anti-monotone de corrélation minimale. Dans le but de se débarrasser des motifs qui sont très rares dans la base *c.-à.-d.* les motifs qui représentent des exceptions et ne sont pas informatifs, une contrainte de fréquence minimale a été également intégrée. L'idée consiste à extraire les N premiers motifs respectant les contraintes de corrélation et de fréquence posées. Ces derniers correspondent aux motifs rares

les plus informatifs.

La problématique d'intégration des contraintes lors de la fouille des motifs corrélés a été aussi étudié dans les travaux proposés dans (Brin et al., 1997) et dans (Grahne et al., 2000). Ces approches traitent de la fouille sous contraintes des motifs corrélés selon le coefficient de corrélation χ^2. Ils exploitent les différentes opportunités d'élagage offertes par ces contraintes et bénéficient du pouvoir sélectif de chaque type de contrainte. Cependant, le coefficient χ^2 ne vérifie pas la contrainte d'anti-monotonie induite par la mesure *bond*. En outre, ces approches se limitent à l'extraction d'un sous-ensemble restreint composé uniquement des motifs minimaux valides *c.-à.-d.* satisfaisant l'ensemble de contraintes posées. De plus, aucune représentation concise des motifs corrélés retenus n'a été proposée.

D'autre part, des approches d'extraction de représentations concises de motifs sous la conjonction de contraintes de types opposés ont été proposées. Par exemple, les auteurs dans (Boulicaut et Jeudy, 2001) ont proposé une approche d'extraction d'une représentation concise basée sur les itemsets libres fréquents et satisfaisant un ensemble donné de contraintes monotones et anti-monotones. L'approche proposée dans (Lei et al., 2003) ainsi que l'algorithme CCI_MINER (Bonchi et Lucchese, 2006) permettent aussi l'extraction sous contraintes d'une représentation concise basée sur les motifs fermés fréquents. Dans ce même cadre s'inscrit aussi l'algorithme CCMINE (Kim et al., 2004) qui permet d'extraire une représentation concise exacte basée sur les motifs fermés fréquents corrélés selon la mesure *all-confidence*. À notre connaissance, ce travail est le premier à s'intéresser aux représentations concises des motifs corrélés rares.

Après avoir cerné les approches d'extraction de motifs sous une conjonction de contraintes de natures différentes, nous introduisons dans ce qui suit les représentations concises exactes des motifs corrélés rares.

4 Nouvelles représentations concises des motifs corrélés rares

Une représentation exacte des motifs corrélés rares doit permettre de déterminer si un motif arbitraire est corrélé rare ou non, et s'il est corrélé rare, la représentation doit permettre de dériver sans perte d'information son support et sa valeur de la mesure *bond*. Dans ce sens, les représentations proposées dans ce travail seront montrées comme étant toujours de taille plus réduite que l'ensemble total des motifs corrélés rares. Elles permettent ainsi une meilleure exploitation des connaissances extraites. En plus, étant sans perte d'information, elles permettent la dérivation quand ceci est nécessaire de tous les motifs corrélés rares non retenus dans une représentation donnée.

Afin de proposer les représentations concises des motifs corrélés rares, nous nous basons sur les classes d'équivalence. Ces classes permettent de ne retenir que les motifs non redondants. En effet, parmi les motifs d'une classe donnée, seuls ceux nécessaire à la régénération de l'ensemble total des motifs corrélés rares seront retenus dans une représentation donnée. Le reste des motifs de la classe ne sera donc pas maintenu, ce qui réduit la redondance dans les connaissances extraites. Ces classes d'équivalence facilitent aussi l'exploration de l'espace de recherche des motifs corrélés rares. En effet, l'application d'un opérateur de fermeture permet de passer de l'élément minimal d'une classe à son élément maximal sans avoir à parcourir les niveaux intermédiaires.

Nous étudions dans un premier temps les spécificités des *classes d'équivalence corrélées rares*, induites par l'opérateur de fermeture f_{bond} et dont les éléments associés sont rares. Sur la base de cette étude, les différentes représentations proposées seront décrites dans la suite.

4.1 Caractérisation des classes d'équivalence corrélées rares

Pour chaque classe d'équivalence induite par l'opérateur de fermeture f_{bond}, les motifs associés admettent la même fermeture par f_{bond}, et sont ainsi caractérisés par les mêmes supports et la même valeur de la mesure *bond* (*cf.* Proposition 2). Par conséquent, les éléments d'une même classe d'équivalence ont le même comportement vis-à-vis des contraintes de corrélation et de rareté. Par conséquent, pour une classe d'équivalence corrélée, *c.-à.-d.* celle contenant des motifs corrélés, les éléments associés sont tous rares ou sont tous fréquents. Il en est d'ailleurs de même pour une classe d'équivalence non corrélée, *c.-à.-d.* dont les motifs associés sont non corrélés. Ainsi, pour une classe induite par f_{bond}, il suffit de tester les contraintes de corrélation et de rareté sur un unique motif de la classe pour avoir l'information concernant tous les autres éléments de la classe.

Nous distinguons en conséquent quatre types de classes : les classes corrélées fréquentes, les classes non corrélées fréquentes, les classes corrélées rares et les classes non corrélées rares. Cette caractéristique des classes d'équivalence induites par f_{bond} est très intéressante. En effet, ceci n'est pas le cas de tous les opérateurs de fermeture. Par exemple, l'application de l'opérateur de fermeture associé au support conjonctif, *c.-à.-d.* l'opérateur de fermeture de Galois (Ganter et Wille, 1999), induit des classes d'équivalence où le comportement d'un motif d'une classe donnée vis-à-vis de la contrainte de corrélation n'est pas représentatif du comportement du reste des motifs de la classe. Pour une classe donnée, chaque motif doit ainsi être testé indépendamment des autres de la même classe pour savoir s'il est corrélé ou non. Il en résulte, de ce qui précède, que l'application de f_{bond} offre un processus sélectif des motifs à retenir comme étant corrélés rares. Ces derniers sont les éléments des classes d'équivalence corrélées rares.

Exemple 14 *Considérons la base $\mathcal{D}$ illustrée par la table 2 pour minsupp = 4 et minbond = 0,2. La figure 2 présente les classes d'équivalence corrélées rares dont les éléments de chacune sont récapitulés comme suit :*

- $\mathcal{C}_1$ *contient le motif* `A`*. Elle admet pour support conjonctif 3 et pour valeur de bond 1.*
- $\mathcal{C}_2$ *contient le motif* `D`*. Elle admet pour support conjonctif 1 et pour valeur de bond 1.*
- $\mathcal{C}_3$ *contient les motifs* `AB`*,* `AE`*,* `ABC`*,* `ABE`*,* `ACE`*, et* `ABCE`*. Elle admet pour support conjonctif 2 et pour valeur de bond* $\frac{2}{5}$*.* `ABCE` *est le fermé corrélé de cette classe.*
- $\mathcal{C}_4$ *contient le motif* `AC`*. Elle admet pour support conjonctif 3 et pour valeur de bond* $\frac{3}{4}$*.*
- $\mathcal{C}_5$ *contient le motif* `AD`*. Elle admet pour support conjonctif 1 et pour valeur de bond* $\frac{1}{3}$*.*
- $\mathcal{C}_6$ *contient les motifs* `CD` *et* `ACD`*. Elle admet pour support conjonctif 1 et pour valeur de bond* $\frac{1}{4}$*.* `ACD` *est le fermé corrélé de cette classe.*
- $\mathcal{C}_7$ *contient les motifs* `BC`*,* `CE` *et* `BCE`*. Elle admet pour support conjonctif 3 et pour valeur de bond* $\frac{3}{5}$*.* `BCE` *est le fermé corrélé de cette classe.*

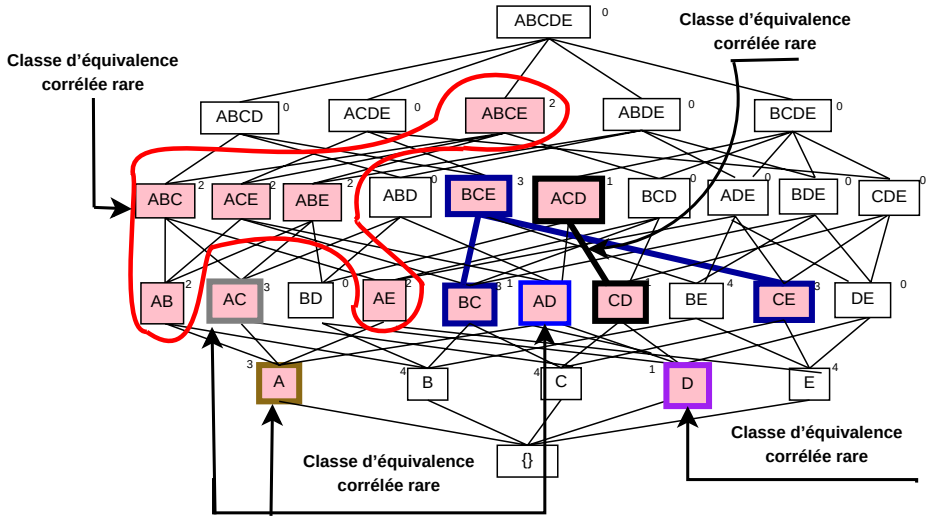

FIG. 2 – *Classes d'équivalence corrélées rares pour minsupp = 4 et minbond = 0,2.*

L'ensemble des motifs corrélés rares est ainsi partitionné en classes disjointes, les classes d'équivalence corrélées rares. Dans chaque classe, un motif fermé corrélé rare est alors le motif le plus large au sens de l'inclusion dans la classe. Par contre, les plus petits motifs sont les motifs minimaux corrélés rares incomparables selon la relation d'inclusion. Les motifs minimaux et fermés seront formellement définis dans ce qui suit.

Définition 17 (***Motifs fermés corrélés rares***) *L'ensemble* $\mathcal{MFCR}$ *des motifs fermés corrélés rares est défini par :* $\mathcal{MFCR} = \{I \in \mathcal{MCR} | \forall\ I_1 \supset I : bond(I) > bond(I_1)\}$

En fait, l'ensemble $\mathcal{MFCR}$ résulte de l'intersection entre l'ensemble des motifs corrélés rares et l'ensemble des motifs fermés corrélés. Ainsi, $\mathcal{MFCR} = \mathcal{MCR} \cap \mathcal{MFC}$.

Définition 18 (***Motifs minimaux corrélés rares***) *L'ensemble* $\mathcal{MMCR}$ *des motifs minimaux corrélés rares est défini par :* $\mathcal{MMCR} = \{I \in \mathcal{MCR} | \forall\ I_1 \subset I : bond(I) < bond(I_1)\}$.

Cet ensemble, $\mathcal{MMCR}$, résulte de l'intersection entre l'ensemble des motifs corrélés rares et l'ensemble des motifs minimaux corrélés. Ainsi, $\mathcal{MMCR} = \mathcal{MCR} \cap \mathcal{MMC}$.

Exemple 15 *Soit la base illustrée par la table 2 pour minsupp = 4 et minbond = 0,2. Le motif* $ACD \in \mathcal{MFCR}$. *En effet, il est fermé corrélé* (*cf. Exemple 9*). *Il est aussi rare* (*cf. Exemple 12*). *Par ailleurs, le motif* $AB \in \mathcal{MMCR}$. *En effet, il est minimal corrélé* (*cf. Exemple 10*). *Il est aussi rare* (*cf. Exemple 12*). *En se référant aux diverses classes d'équivalence corrélées rares présentées dans la figure 2, nous avons l'ensemble* $\mathcal{MFCR}$ *composé des éléments maximaux de ces classes c.-à.-d.* A, D, AC, AD, ACD, BCE *et* $ABCE$. *Par ailleurs, l'ensemble* $\mathcal{MMCR}$ *est composé des éléments minimaux de ces classes c.-à.-d.* A, D, AB, AC, AD, AE, BC, CD *et* CE. *Comme le montre la figure 2, les motifs* A, D, AC *et* AD *sont à la fois des motifs fermés et minimaux. Leurs classes associées se réduisent donc chacune à un unique élément.*

Après avoir détaillé les propriétés des classes d'équivalence corrélées rares, nous introduisons dans la suite les représentations concises proposées.

4.2 Représentations concises exactes des motifs corrélés rares

Une première idée intuitive afin de proposer une représentation concise exacte des motifs corrélés rares serait de voir si les éléments minimaux ou les éléments maximaux des classes d'équivalence associées permettraient de représenter sans perte d'information cet ensemble. Dans ce sens, il est important de rappeler que l'ensemble $\mathcal{MCR}$ des motifs corrélés rares résulte de l'intersection de l'idéal d'ordre des motifs corrélés et du filtre d'ordre des motifs rares. Cet ensemble $\mathcal{MCR}$ ne forme donc ni un idéal d'ordre ni un filtre d'ordre. Dans cette situation, pris chacun indépendamment de l'autre, est ce que l'ensemble $\mathcal{MMCR}$ ou l'ensemble $\mathcal{MFCR}$ peut constituer une représentation concise exacte des motifs rares corrélés ?

Analysons dans ce qui suit chacun de ces deux ensembles :

- Commençons par l'ensemble $\mathcal{MMCR}$ des minimaux des classes d'équivalence corrélées rares. De part la nature de ses éléments – minimaux de leurs classes d'équivalence – cet ensemble permet pour un motif donné I de vérifier s'il est rare ou non. En effet, il suffit de trouver un élément $J \in \mathcal{MMCR}$, tel que $J \subseteq I$ pour décider que I est un motif rare. Si ce n'est pas le cas, alors I n'est pas un motif rare. Toutefois, l'ensemble $\mathcal{MMCR}$ ne permet pas de déterminer dans le cas général si I est corrélé ou non (ceci n'est possible que si $I \in \mathcal{MMCR}$). En effet, même s'il existe $J \in \mathcal{MMCR}$, tel que $J \subset I$, et même sachant que J est corrélé, nous ne pouvons rien décider quant au statut de I vis-à-vis de la contrainte de corrélation puisque cette dernière est anti-monotone (le fait que J est corrélé n'implique pas que I l'est aussi). Ainsi, $\mathcal{MMCR}$ ne peut pas constituer une représentation exacte de $\mathcal{MCR}$.

- Traitons maintenant le cas de l'ensemble $\mathcal{MFCR}$ des maximaux des classes d'équivalence corrélées rares. D'une manière duale à $\mathcal{MMCR}$, les éléments de $\mathcal{MFCR}$ permettent de déterminer pour un motif I s'il est corrélé ou non. Il suffit qu'il soit inclus dans un motif $J \in \mathcal{MFCR}$, et sinon I n'est pas corrélé. Toutefois, de part leur nature, les fermés appartenant à $\mathcal{MFCR}$ ne permettent pas dans le cas général de dériver l'information concernant le statut de rareté d'un motif I quelconque (ceci n'est possible que si $I \in \mathcal{MFCR}$). En effet, même s'il existe $J \in \mathcal{MFCR}$, tel que $I \subset J$, et même sachant que J est rare, nous ne pouvons pas savoir si I est rare ou non puisque la contrainte de rareté est monotone (le fait que J est rare n'implique pas que I l'est aussi). Ainsi, $\mathcal{MFCR}$ ne peut pas constituer une représentation exacte de $\mathcal{MCR}$.

Il résulte de l'analyse que nous venons d'effectuer que la complémentarité entre $\mathcal{MMCR}$ et $\mathcal{MFCR}$ peut constituer une représentation exacte des motifs corrélés rares. Cette première alternative est étudiée dans la sous-section qui suit, et qui sera suivie par deux optimisations afin de ne retenir que les éléments indispensables à la régénération sans perte d'information des éléments de $\mathcal{MCR}$.

4.2.1 La représentation concise exacte $\mathcal{RMCR}$

La première représentation que nous proposons est définie comme suit.

Définition 19 (***Représentation*** $\mathcal{RMCR}$) *Soit $\mathcal{RMCR}$ la représentation concise exacte des motifs corrélés rares basée sur l'ensemble $\mathcal{MFCR}$ des motifs fermés corrélés rares et sur*

l'ensemble $\mathcal{MMCR}$ des motifs minimaux corrélés rares. La représentation $\mathcal{RMCR}$ est définie comme suit : $\mathcal{RMCR}$ = $\mathcal{MFCR} \cup \mathcal{MMCR}$. Chaque élément I de $\mathcal{RMCR}$ est muni de son support, Supp($\wedge I$), et sa mesure bond, bond(I).

Exemple 16 *Considérons la base de transactions donnée dans la table 2, pour minsupp = 4 et minbond = 0,2. En considérant les ensembles $\mathcal{MFCR}$ et $\mathcal{MMCR}$ (cf. Exemple 15), la représentation $\mathcal{RMCR}$ est composée par :* (A, 3, $\frac{3}{3}$), (D, 1, $\frac{1}{1}$), (AB, 2, $\frac{2}{5}$), (AC, 3, $\frac{3}{4}$), (AD, 1, $\frac{1}{3}$), (AE, 2, $\frac{2}{5}$), (BC, 3, $\frac{3}{5}$), (CD, 1, $\frac{1}{4}$), (CE, 3, $\frac{3}{5}$), (ACD, 1, $\frac{1}{4}$), (BCE, 3, $\frac{3}{5}$) *et* (ABCE, 2, $\frac{2}{5}$).

Le théorème suivant montre que les éléments de $\mathcal{RMCR}$ représentent sans perte d'information les motifs corrélés rares.

Théorème 1 *La représentation $\mathcal{RMCR}$ est une représentation concise exacte de l'ensemble $\mathcal{MCR}$ des motifs corrélés rares.*

Preuve. Soit un motif $I \subseteq \mathcal{I}$. Trois cas se présentent :

a) Si $I \in \mathcal{RMCR}$, alors I est un motif corrélé rare et nous avons son support et sa valeur de la mesure *bond*.

b) Si $\nexists\ J \in \mathcal{RMCR}$ tel que $J \subseteq I$ ou $\nexists\ Z \in \mathcal{RMCR}$ tel que $I \subseteq Z$, alors $I \notin \mathcal{MCR}$ puisque I n'appartient à aucune classe d'équivalence corrélée rare.

c) Sinon, $I \in \mathcal{MCR}$. En effet, d'après la proposition 3, I est corrélé puisque inclus dans un motif corrélé, à savoir Z. Il est aussi rare puisque englobant un motif rare, à savoir J. Dans ce cas, il suffit de localiser la fermeture de I par f_{bond}, disons F. F appartient nécessairement à $\mathcal{RMCR}$ puisque I est un motif corrélé rare et $\mathcal{RMCR}$ inclut l'ensemble $\mathcal{MFCR}$ des motifs fermés corrélés rares. Par conséquent, $F = min_{\subseteq}\{I_1 \in \mathcal{RMCR} |\ I \subseteq I_1\}$. Comme l'opérateur f_{bond} préserve la mesure *bond* et par conséquent le support conjonctif (*cf.* Proposition 2), nous avons : *bond*(I) = *bond*(F) et *Supp*($\wedge I$) = *Supp*($\wedge F$). $\diamondsuit$

Exemple 17 *Considérons la représentation $\mathcal{RMCR}$ donnée dans l'exemple précédent. Illustrons chacun des trois cas. Le motif AD $\in \mathcal{RMCR}$. Ainsi, nous avons son support égal à 1 et sa valeur de la mesure bond égale à $\frac{1}{3}$. Considérons le motif BE. Bien qu'il soit inclus dans deux motifs de $\mathcal{RMCR}$, à savoir BCE et ABCE, BE $\notin \mathcal{MCR}$ puisque aucun élément de $\mathcal{RMCR}$ n'est inclus dans BE. Soit maintenant le motif ABC. Il existe deux motifs de $\mathcal{RMCR}$ qui vérifient la condition faisant de ABC un motif corrélé rare, à savoir AB et ABCE, puisque AB $\subseteq$ ABC $\subseteq$ ABCE. Le plus petit motif de $\mathcal{RMCR}$ couvrant ABC, c.-à.-d. sa fermeture, est ABCE. Ainsi, bond(ABC) = bond(ABCE) = $\frac{2}{5}$, et Supp($\wedge$ABC) = Supp($\wedge$ABCE) = 2.*

Il est important de noter que la représentation $\mathcal{RMCR}$ est une *couverture parfaite* de l'ensemble $\mathcal{MCR}$. En effet, la taille de la représentation $\mathcal{RMCR}$ *ne dépasse jamais* celle de l'ensemble $\mathcal{MCR}$ quelle que soit la base et les valeurs de *minsupp* et de *minbond* considérées. En effet, nous avons toujours ($\mathcal{MFCR} \cup \mathcal{MMCR}$) $\subseteq \mathcal{MCR}$.

Par ailleurs, connaissant le support conjonctif d'un motif et sa valeur de la mesure *bond*, nous pouvons calculer son support disjonctif et par conséquent son support négatif. L'interrogation de la représentation $\mathcal{RMCR}$ peut donc se baser sur la preuve du théorème 1. Ainsi,

pour un motif arbitraire, ayant la représentation $\mathcal{RMCR}$, l'utilisateur peut savoir s'il est corrélé rare ou non. S'il est corrélé rare, les informations sur les différents supports et la mesure *bond* qui lui sont associées seront dérivées en utilisant le mécanisme indiqué dans la preuve susmentionnée.

Le processus de régénération de l'ensemble total des motifs corrélés rares peut aussi se baser sur cette preuve. Ce processus commence par les plus petits motifs corrélés rares à savoir les motifs minimaux corrélés rares (formant l'ensemble $\mathcal{MMCR}$). Ces motifs appartiennent à $\mathcal{RMCR}$ et nous avons donc les informations les concernant. Il suffira après de localiser pour chaque minimal M sa fermeture F et qui se trouve nécessairement dans $\mathcal{RMCR}$ ($F \in \mathcal{MFCR}$ et cet ensemble est contenu dans $\mathcal{RMCR}$). Tous les motifs compris entre M et F admettent les mêmes informations que M et F puisqu'ils appartiennent à la même classe d'équivalence corrélée rare.

Remarque 1 *Il est important de noter que nous sommes obligés de maintenir, pour un motif I de la représentation, à la fois le Supp($\wedge I$) et bond(I). D'une part, la valeur de bond(I) étant un rapport entre le support conjonctif et celui disjonctif de I ne permet pas de dériver le support conjonctif de I. D'autre part, ayant le support conjonctif d'un motif I, ceci n'est pas suffisant pour calculer la valeur de sa mesure bond. En effet, ceci nécessite la connaissance de son support disjonctif. Ce dernier ne peut être dérivé moyennant les identités d'inclusion-exclusion que connaissant les supports conjonctifs de tous les sous-ensembles de I (Galambos et Simonelli, 2000). Toutefois, si I est un motif corrélé rare, tous ses sous-ensembles ne le sont pas forcément. Par conséquent, nous n'avons pas accès à leurs supports conjonctifs respectifs. Ainsi, il faut retenir le support conjonctif et la valeur de la mesure bond pour chaque élément de la représentation. C'est aussi la raison pour laquelle les représentations concises basées sur les règles de déduction (Calders et al., 2005) et celles basées sur les identités d'inclusion-exclusion (Casali et al., 2005; Hamrouni et al., 2009) ne sont pas applicables pour représenter l'ensemble des motifs corrélés rares. En effet, ces représentations nécessitent pour un motif donné la connaissance du support conjonctif ou disjonctif, suivant la représentation, associé à tous ses sous-ensembles. Ceci est exigé que ce soit pour retenir les éléments de la représentation ou pour la dérivation des informations associées aux éléments non retenus dans la représentation.*

Dans cette situation, les motifs fermés et minimaux des classes d'équivalence offrent comme montré précédemment une solution intéressante pour représenter d'une manière concise l'ensemble des motifs corrélés rares. En effet, la localisation de tels motifs nécessite un voisinage restreint, les sur-ensembles immédiats et les sous-ensembles immédiats respectivement, et non tous leurs sous-ensembles respectifs, comme c'est le cas par exemple des motifs non dérivables (Calders et Goethals, 2007). En plus, la dérivation des supports des motifs à partir des fermés et des minimaux est réalisée d'une manière directe, contrairement par exemple aux motifs essentiels (Casali et al., 2005) et aux motifs non dérivables qui nécessitent tous les sous-ensembles du motif dérivé.

Remarque 2 *Il est aussi intéressant de signaler que le fait de considérer, dans $\mathcal{RMCR}$, l'union entre les ensembles $\mathcal{MMCR}$ et $\mathcal{MFCR}$ permet d'éviter la redondance – à cause de la duplication d'un motif donné – qui peut apparaître dans la représentation si nous considérons chacun des ensembles $\mathcal{MMCR}$ et $\mathcal{MFCR}$ séparément. Par exemple, si nous considérons l'exemple 16, nous remarquons que les éléments* (A, 3, $\frac{3}{3}$), (D, 1, $\frac{1}{1}$), (AC, 3, $\frac{3}{4}$) *et* (AD, 1, $\frac{1}{3}$)

appartiennent à la fois à l'ensemble $\mathcal{MMCR}$ et $\mathcal{MFCR}$. Toutefois, un avantage de la gestion de chaque ensemble à part est la réduction de certains tests d'inclusion lors de l'interrogation de la représentation. Le choix entre tolérer une certaine duplication et réduire éventuellement le coût de la régénération dépend de la nature de l'application où il sera éventuellement question de privilégier soit l'espace mémoire soit les temps de dérivation. Notons qu'une solution intermédiaire serait de localiser dans un premier groupe les éléments qui sont à la fois des motifs fermés et des motifs minimaux tels que A, D, AC *et* AD *dans notre cas, dans un second le reste des minimaux à part, et le reste des fermés formeront un troisième groupe. Le premier et le second groupe seront utilisés pour les traitements où les minimaux seront utiles, tandis que le premier et le troisième seront utilisés pour les traitements nécessitant les fermés.*

Nous proposons dans la suite de cette section deux optimisations de $\mathcal{RMCR}$ permettant de réduire encore plus le nombre de motifs à retenir dans la représentation tout en garantissant la non-perte d'information concernant l'ensemble $\mathcal{MCR}$ des motifs corrélés rares.

4.2.2 La représentation concise exacte $\mathcal{RMMaxF}$

Cette première optimisation se base sur le fait que l'ensemble $\mathcal{MMCR}$ des motifs minimaux corrélés rares augmenté seulement des maximaux par rapport à l'inclusion ensembliste, parmi les motifs fermés corrélés rares, est suffisant pour représenter d'une manière exacte l'ensemble $\mathcal{MCR}$. L'ensemble $\mathcal{MFCRMax}$ des motifs fermés corrélés rares maximaux est défini comme suit :

Définition 20 (***Ensemble*** $\mathcal{MFCRMax}$ ***des motifs fermés corrélés rares maximaux***) *L'ensemble $\mathcal{MFCRMax}$ correspond aux motifs qui sont à la fois des motifs fermés corrélés rares* (*cf. Définition 17, page 16*) *et des motifs corrélés maximaux* (*cf. Définition 11, page 8*). *Ainsi, $\mathcal{MFCRMax} = \mathcal{MFCR} \cap \mathcal{MCMax}$.*

L'ensemble $\mathcal{MFCRMax}$ est donc restreint aux éléments de $\mathcal{MCMax}$ qui sont aussi rares (en plus d'être les plus grands motifs corrélés).

Exemple 18 *Soit la base illustrée par la table 2. Pour minsupp = 4 et minbond = 0,2, $\mathcal{MFCR}$ = {*A, D, AC, AD, ACD, BCE, ABCE*} (cf. Exemple 15). Par ailleurs, pour minbond = 0,2, $\mathcal{MCMax}$ = {*ACD, ABCE*} (cf. Exemple 7). Ainsi, $\mathcal{MFCRMax} = \mathcal{MFCR} \cap \mathcal{MCMax}$ = {*ACD, ABCE*}. En effet, les fermés corrélés* A, D, AC *et* AD *ne sont pas retenus puisqu'ils sont inclus dans* ACD. *Le fermé* BCE *est aussi éliminé puisqu'il est inclus dans* ABCE.

La définition suivante présente la représentation des motifs corrélés rares basée sur cette optimisation.

Définition 21 (***Représentation*** $\mathcal{RMMaxF}$) *Soit $\mathcal{RMMaxF}$ la représentation basée sur l'ensemble $\mathcal{MMCR}$ et l'ensemble $\mathcal{MFCRMax}$. Nous avons $\mathcal{RMMaxF} = \mathcal{MMCR} \cup \mathcal{MFCRMax}$. Chaque élément I de $\mathcal{RMMaxF}$ est muni de son support, Supp($\wedge I$), et sa mesure bond, bond(I).*

Exemple 19 *Considérons la base de transactions donnée dans la table 2, Pour minsupp = 4 et minbond = 0,2. La représentation $\mathcal{RMMaxF}$ est composée par :* (A, 3, $\frac{3}{3}$), (D, 1, $\frac{1}{1}$), (AB,

2, $\frac{2}{5}$), (AC, 3, $\frac{3}{4}$), (AD, 1, $\frac{1}{3}$), (AE, 2, $\frac{2}{5}$), (BC, 3, $\frac{3}{5}$), (CD, 1, $\frac{1}{4}$), (CE, 3, $\frac{3}{5}$), (ACD, 1, $\frac{1}{4}$), *et* (ABCE, 2, $\frac{2}{5}$). *Nous remarquons que, pour cet exemple, le seul élément appartenant à la représentation $\mathcal{RMCR}$ et non à la représentation $\mathcal{RMM}ax\mathcal{F}$ est le motif* BCE*. En effet, ceci est dû au fait que les motifs fermés éliminés de $\mathcal{MCM}ax$ sont eux mêmes des minimaux corrélés rares, à savoir* A*,* D*,* AC *et* AD*. Toutefois, la représentation $\mathcal{RMM}ax\mathcal{F}$ serait plus réduite que $\mathcal{RMCR}$ si les ensembles $\mathcal{MMCR}$ et $\mathcal{MFCR}$ sont gérés séparément* (*cf. Remarque 2*). *En effet, il n'y aura plus de duplication de* A*,* D*,* AC *et de* AD*.*

Le théorème 2 montre que $\mathcal{RMM}$ax$\mathcal{F}$ couvre sans perte d'information $\mathcal{MCR}$.

Théorème 2 *La représentation $\mathcal{RMM}ax\mathcal{F}$ est une représentation concise exacte de l'ensemble $\mathcal{MCR}$ des motifs corrélés rares.*

Preuve. Soit un motif $I \subseteq \mathcal{I}$. Trois cas se présentent :

a) Si $I \in \mathcal{RMM}ax\mathcal{F}$, alors I est un motif corrélé rare et nous avons son support et sa valeur de la mesure *bond*.

b) Si $\nexists\ J \in \mathcal{RMM}ax\mathcal{F}$ tel que $J \subseteq I$ ou $\nexists\ Z \in \mathcal{RMM}ax\mathcal{F}$ tel que $I \subseteq Z$, alors $I \notin \mathcal{MCR}$ puisque I n'appartient à aucune classe d'équivalence corrélée rare.

c) Sinon, $I \in \mathcal{MCR}$. En effet, d'après la proposition 3, I est corrélé puisque inclus dans un motif corrélé, à savoir Z. Il est aussi rare puisque englobant un motif rare, à savoir J. Comme I est un motif corrélé rare et la représentation $\mathcal{RMM}ax\mathcal{F}$ inclut l'ensemble $\mathcal{MMCR}$ contenant les éléments minimaux des différentes classes d'équivalence corrélées rares, cette représentation contient au moins un élément de la classe d'équivalence de I, en particulier tous les motifs minimaux de la classe.
Comme le support conjonctif et la mesure *bond* décroissent avec la taille des motifs, les valeurs du support conjonctif et de la mesure *bond* de I sont égales aux valeurs minimales des mesures associées à ses sous-ensembles appartenant à $\mathcal{RMM}ax\mathcal{F}$. Il en résulte que :

- *Supp*($\wedge I$) = $min\{$*Supp*($\wedge I_1$)$|\ I_1 \in \mathcal{RMM}ax\mathcal{F}$ et $I_1 \subseteq I\}$, et,
- *bond*(I) = $min\{$*bond*(I_1)$|\ I_1 \in \mathcal{RMM}ax\mathcal{F}$ et $I_1 \subseteq I\}$. $\diamondsuit$

Exemple 20 *Soit la représentation $\mathcal{RMM}ax\mathcal{F}$ donnée dans l'exemple précédent. Le traitement du premier et du second cas est semblable à ceux des deux premiers cas de la représentation $\mathcal{RMCR}$* (*cf. Exemple 17*). *Considérons donc le motif* ABE *pour illustrer le troisième cas. Il existe deux motifs de $\mathcal{RMM}ax\mathcal{F}$ qui vérifient la condition faisant de* ABE *un motif corrélé rare, à savoir* AB *et* ABCE (AB $\subseteq$ ABE $\subseteq$ ABCE). *Les motifs de $\mathcal{RMM}ax\mathcal{F}$ inclus dans* ABE *sont* AB *et* AE*. Par conséquent, Supp*($\wedge$ABE) = $min\{$*Supp*($\wedge$*AB*), *Supp*($\wedge$*AE*)$\}$ = $min\{2, 2\}$ = 2*, et bond*(ABE) = $min\{$*bond*(*AB*), *bond*(*AE*)$\}$ = $min\{\frac{2}{5}, \frac{2}{5}\} = \frac{2}{5}$.

Étant incluse dans $\mathcal{RMCR}$, qui a été montrée comme étant une couverture parfaite de $\mathcal{MCR}$, la représentation $\mathcal{RMM}ax\mathcal{F}$ est aussi une couverture parfaite de $\mathcal{MCR}$.

La sous-section suivante présente une autre optimisation de la représentation $\mathcal{RMCR}$.

4.2.3 La représentation concise exacte $\mathcal{RM}in\mathcal{MF}$

D'une manière duale à la représentation précédente, il suffit de retenir dans $\mathcal{RMCR}$ que les motifs minimaux, par rapport à l'inclusion ensembliste, parmi ceux de l'ensemble $\mathcal{MMCR}$

des motifs minimaux corrélés rares. L'élagage des autres éléments de $\mathcal{MMCR}$ sera prouvé comme étant sans perte d'information lors de la régénération de l'ensemble $\mathcal{MCR}$ des motifs corrélés rares. L'ensemble $\mathcal{MMCRM}in$ des motifs minimaux parmi ceux de $\mathcal{MMCR}$ est défini comme suit :

Définition 22 (***Ensemble*** $\mathcal{MMCRM}in$ ***des éléments minimaux de l'ensemble*** $\mathcal{MMCR}$) *L'ensemble* $\mathcal{MMCRM}in$ *contient les motifs qui sont à la fois des motifs minimaux corrélés rares* (*cf. Définition 18, page 16*) *et des motifs rares minimaux* (*cf. Définition 8, page 7*). *Ainsi,* $\mathcal{MMCRM}in = \mathcal{MMCR} \cap \mathcal{MRM}in$.

L'ensemble $\mathcal{MMCRM}in$ est donc restreint aux éléments de $\mathcal{MRM}in$ qui sont aussi corrélés (en plus d'être les plus petits motifs rares).

Exemple 21 *Soit la base illustrée par la table 2. Pour minsupp = 4 et minbond = 0,2,* $\mathcal{MMCR}$ = {A, D, AB, AC, AD, AE, BC, CD, CE} (*cf. Exemple 15*). *Par ailleurs, pour minsupp = 4,* $\mathcal{MRM}in$ = {A, D, BC, CE}. *Ainsi,* $\mathcal{MMCRM}in = \mathcal{MMCR} \cap \mathcal{MRM}in$ = {A, D, BC, CE}. *Les motifs minimaux corrélés rares* AB, AC, AD *et* AE *ne sont pas retenus puisqu'ils englobent l'item rare* A. *De même pour le motif* CD, *il est aussi éliminé puisqu'il englobe l'item rare* D.

Remarque 3 *Il est important de noter que dans l'exemple précédent, nous avons* $\mathcal{MRM}in \subseteq \mathcal{MMCR}$. *Toutefois, ceci n'est pas le cas d'une manière générale. En effet, un motif rare minimal peut bien évidemment ne pas être corrélé et donc ne pas appartenir à* $\mathcal{MMCR}$. *Cette remarque s'applique aussi pour le cas de l'exemple 18 où* $\mathcal{MCM}ax \subseteq \mathcal{MFCR}$. *En effet, un motif corrélé maximal peut ne pas être rare et donc ne pas appartenir à* $\mathcal{MFCR}$.

La définition 23 présente la représentation résultante de l'utilisation de $\mathcal{MMCRM}in$.

Définition 23 (***Représentation*** $\mathcal{RM}in\mathcal{MF}$) *Soit* $\mathcal{RM}in\mathcal{MF}$ *la représentation basée sur l'ensemble* $\mathcal{MFCR}$ *et l'ensemble* $\mathcal{MMCRM}in$. *Nous avons* $\mathcal{RM}in\mathcal{MF} = \mathcal{MFCR} \cup \mathcal{MMCRM}in$. *Chaque élément* I *de* $\mathcal{RM}in\mathcal{MF}$ *est muni de son support, Supp*($\wedge I$), *et sa mesure bond, bond*(I).

Exemple 22 *Considérons la base de transactions donnée dans la table 2, pour minsupp = 4 et minbond = 0,2. La représentation* $\mathcal{RM}in\mathcal{MF}$ *est composée par :* (A, 3, $\frac{3}{3}$), (D, 1, $\frac{1}{1}$), (AC, 3, $\frac{3}{4}$), (AD, 1, $\frac{1}{3}$), (BC, 3, $\frac{3}{5}$), (CE, 3, $\frac{3}{5}$), (ACD, 1, $\frac{1}{4}$), (BCE, 3, $\frac{3}{5}$), *et* (ABCE, 2, $\frac{2}{5}$). *Nous remarquons que, comparée à* $\mathcal{RMCR}$, *cette représentation admet trois éléments en moins à savoir* AB, AE *et* CD

Le théorème prouve que cette troisième représentation est aussi sans perte d'information de l'ensemble $\mathcal{MCR}$ des motifs corrélés rares.

Théorème 3 *La représentation* $\mathcal{RM}in\mathcal{MF}$ *est une représentation concise exacte de l'ensemble* $\mathcal{MCR}$ *des motifs corrélés rares.*

Preuve. Soit un motif $I \subseteq \mathcal{I}$. Trois cas se présentent :

a) Si $I \in \mathcal{RM}in\mathcal{MF}$, alors I est un motif corrélé rare et nous avons son support et sa valeur de la mesure *bond*.

b) Si $\nexists\ J \in \mathcal{RM}in\mathcal{MF}$ tel que $J \subseteq I$ ou $\nexists\ Z \in \mathcal{RM}in\mathcal{MF}$ tel que $I \subseteq Z$, alors $I \notin \mathcal{MCR}$ puisque I n'appartient à aucune classe d'équivalence corrélée rare.

c) Sinon, $I \in \mathcal{MCR}$. En effet, d'après la proposition 3, I est corrélé puisque inclus dans un motif corrélé, à savoir Z. Il est aussi rare puisque englobant un motif rare, à savoir J. Comme l'ensemble $\mathcal{MFCR}$ appartient à $\mathcal{RM}in\mathcal{MF}$, il suffit de localiser le fermé corrélé de I, disons F, égal à : $F = min_{\subseteq}\{I_1 \in \mathcal{RM}in\mathcal{MF} \mid I \subseteq I_1\}$. Ainsi, *bond*($I$) = *bond*($F$) et *Supp*($\wedge I$) = *Supp*($\wedge F$). $\diamondsuit$

Prière de se référer à l'exemple 17 (page 18) pour un exemple de déroulement les traitements associés à cette représentation qui sont similaires à ceux de la représentation $\mathcal{RMCR}$.

Étant incluse dans $\mathcal{RMCR}$, $\mathcal{RM}in\mathcal{MF}$ est comme les deux précédentes représentations une couverture parfaite de $\mathcal{MCR}$.

Remarque 4 *La représentation $\mathcal{RM}in\mathcal{MM}ax\mathcal{F}$ = $\mathcal{MFCRM}ax \cup \mathcal{MMCRM}in$ n'est qu'une représentation approximative de $\mathcal{MCR}$. En effet, pour un motif arbitraire $I \subseteq \mathcal{I}$, cette représentation permet de déterminer si I est corrélé rare ou non. Il suffit de trouver deux motifs J et Z appartenant à la représentation tel que $J \subseteq I \subseteq Z$. Si J ou Z n'existe pas alors $I \notin \mathcal{MCR}$. Toutefois, les informations concernant le support et la mesure bond de I ne peuvent être exactement dérivées que si $I \in \mathcal{RM}in\mathcal{MM}ax\mathcal{F}$. Dans le cas contraire, cette représentation ne permet pas de les dériver d'une manière exacte étant donné qu'elle peut ne contenir aucun élément représentatif de la classe d'équivalence de I (c.-à.-d. ni le fermé associé s'il n'appartient pas à $\mathcal{MFCRM}ax$ ni les minimaux associés s'il n'appartiennent pas à $\mathcal{MMCRM}in$). Seule une approximation des supports de I et de sa valeur de bond peut être effectuée dans ce cas.*

5 Conclusion et perspectives

Dans cet article, nous nous sommes intéressés à la fouille des motifs corrélés rares associés à la mesure *bond*. Ces motifs résultent de la conjonction de deux contraintes de types opposés, à savoir la contrainte anti-monotone de la corrélation et la contrainte monotone de la rareté. Cette nature opposée des contraintes traitées dans ce travail rend complexe la localisation de l'ensemble des motifs corrélés rares. À cet égard, la notion de bordure a été utilisée afin de délimiter l'espace associé dans le treillis des motifs. En utilisant l'opérateur de fermeture associé à la mesure *bond*, l'ensemble des motifs corrélés rares a été partitionné en groupes disjoints, les classes d'équivalence corrélées rares. Chaque classe regroupe les motifs ayant les mêmes caractéristiques. Dans ce sens, de nouvelles représentations concises des motifs corrélés rares ont été proposées en utilisant les éléments maximaux et ceux minimaux des classes d'équivalence.

Les perspectives de travaux futurs concernent : (i) la proposition d'un algorithme optimisé dédié à la fouille des représentations proposées ainsi qu'un mécanisme efficace de régénération des motifs corrélés rares. À cette fin, le cadre générique proposé dans (Bucila et al., 2003) ainsi que les récentes optimisations proposées dans la littérature (Segond et Borgelt, 2011) peuvent être adaptées ; (ii) la réalisation d'une étude expérimentale de l'apport en terme de compacité

et de temps d'extraction des représentations proposées. À cet égard, la génération des règles d'association de classification à partir de ces représentations et leurs applications dans la détection d'intrusions semble être une piste prometteuse. La fouille des motifs rares à cette fin s'avère en effet intéressante puisque les motifs rares permettent de repérer les évènements rares survenus et dont une partie consiste en des tentatives d'intrusion. Dans cette situation, les représentations proposées offrent l'information concernant les différents supports d'un motif en plus de sa mesure *bond*. Ceci permet d'extraire et d'appliquer dans des cas réels des formes généralisées de règles d'association présentant des conjonctions, des disjonctions, et des négations d'items corrélés rares en prémisse ou en conclusion ; et (iii) l'extension de l'approche proposée dans ce travail pour les motifs corrélés rares selon toute autre mesure de corrélation vérifiant les mêmes propriétés que la mesure *bond*, la mesure *all-confidence* (Omiecinski, 2003) par exemple. L'évaluation sur la base d'une application réelle de la qualité des motifs corrélés rares associés à chacune de ces mesures est alors une perspective intéressante.

Références

Agrawal, R. et R. Srikant (1994). Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases (VLDB 1994), Santiago, Chile*, pp. 487–499.

Ben Younes, N., T. Hamrouni, et S. Ben Yahia (2010). Bridging conjunctive and disjunctive search spaces for mining a new concise and exact representation of correlated patterns. In *Proceedings of the 13th International Conference Discovery Science (DS 2010), LNCS, volume 6332, Springer-Verlag, Canberra, Australia*, pp. 189–204.

Berberidis, C. et I. P. Vlahavas (2007). Detection and prediction of rare events in transaction databases. *International Journal on Artificial Intelligence Tools 16(5)*, 829–848.

Boley, M. et T. Gärtner (2009). On the complexity of constraint-based theory extraction. In *Proceedings of the 12th International Conference Discovery Science (DS 2009), LNCS, volume 5808, Springer-Verlag, Porto, Portugal*, pp. 92–106.

Bonchi, F., F. Giannotti, A. Mazzanti, et D. Pedreschi (2005). Efficient breadth-first mining of frequent pattern with monotone constraints. *Knowledge and Information Systems 8(2)*, 131–153.

Bonchi, F. et C. Lucchese (2006). On condensed representations of constrained frequent patterns. *Knowledge and Information Systems 9(2)*, 180–201.

Booker, Q. E. (2009). Improving identity resolution in criminal justice data : An application of NORA and SUDA. *Journal of Information Assurance and Security 4*, 403–411.

Boulicaut, J.-F. et B. Jeudy (2001). Mining free itemsets under constraints. In *Proceedings of the 5th International Database Engineering & Applications Symposium (IDEAS 01), IEEE Computer Society Press, Grenoble, France*, pp. 322–329.

Boulicaut, J.-F. et B. Jeudy (2010). Constraint-based data mining. In *Data Mining and Knowledge Discovery Handbook (2nd edition), Springer*, pp. 339–354.

Brin, S., R. Motwani, et C. Silverstein (1997). Beyond market baskets : generalizing association rules to correlations. In *Proceedings of the ACM-SIGMOD International Conference on Management of Data (SIGMOD 1997), Washington D. C., USA*, pp. 265–276.

Bucila, C., J. Gehrke, D. Kifer, et W. M. White (2003). DUALMINER : A dual-pruning algorithm for itemsets with constraints. *Data Mining knowledge Discovery 7(3)*, 241–272.

Calders, T. et B. Goethals (2007). Non-derivable itemset mining. *Data Mining and Knowledge Discovery 14(1)*, 171–206.

Calders, T., C. Rigotti, et J.-F. Boulicaut (2005). A survey on condensed representations for frequent sets. In *Constraint Based Mining and Inductive Databases, LNAI, volume 3848, Springer-Verlag*, pp. 64–80.

Casali, A., R. Cicchetti, et L. Lakhal (2005). Essential patterns : A perfect cover of frequent patterns. In *Proceedings of the 7th International Conference on Data Warehousing and Knowledge Discovery (DaWaK 2005), LNCS, volume 3589, Springer-Verlag, Copenhagen, Denmark*, pp. 428–437.

Cohen, E., M. Datar, S. Fujiwara, A. Gionis, P. Indyk, R.Motwani, J. D. Ullman, et C. Yang (2000). Finding interesting associations without support pruning. In *Proceedings of the 16th International Conference on Data Engineering (ICDE 2000), IEEE Computer Society Press, San Diego, California, USA*, pp. 489–499.

De Raedt, L., M. Jaeger, S. D. Lee, et H. Mannila (2002). A theory of inductive query answering. In *Proceedings of the 2nd International Conference on Data Mining (ICDM 2002), IEEE Computer Society Press, Maebashi City, Japan*, pp. 123–130.

El-Hajj, M. et O. R. Zaïane (2005). Mining with constraints by pruning and avoiding ineffectual processing. In *Proceedings of the 18th Australian Joint Conference on Artificial Intelligence (AI 2005), LNCS, volume 3809, Springer-Verlag, Sydney, Australia*, pp. 1001–1004.

El-Hajj, M., O. R. Zaïane, et P. Nalos (2005). Bifold constraint-based mining by simultaneous monotone and anti-monotone checking. In *Proceedings of the 5th International Conference on Data Mining (ICDM 2005), IEEE Computer Society Press, Houston, Texas, USA*, pp. 146–153.

Galambos, J. et I. Simonelli (2000). *Bonferroni-type inequalities with applications*. Springer.

Ganter, B. et R. Wille (1999). *Formal Concept Analysis*. Springer.

Grahne, G., L. V. S. Lakshmanan, et X. Wang (2000). Efficient mining of constrained correlated sets. In *Proceedings of the 16th International Conference on Data Engineering (ICDE 2000), IEEE Computer Society Press, San Diego, California, USA*, pp. 512–521.

Hamrouni, T., S. Ben Chaabene, et S. Ben Yahia (2011). Réduire pour mieux exploiter : représentations concises et exactes des motifs rares. In *Actes du 7ième Atelier Qualité des Données et des Connaissances, en conjonction avec la 11ième Conférence Internationale Francophone Extraction et Gestion des Connaissances (EGC 2011), 25 - 28 Janvier, Brest, France*, pp. 1–12.

Hamrouni, T., S. Ben Yahia, et E. Mephu Nguifo (2009). Sweeping the disjunctive search space towards mining new exact concise representations of frequent itemsets. *Data & Knowledge Engineering 68(10)*, 1091–1111.

Hamrouni, T., S. Ben Yahia, et E. Mephu Nguifo (2010). Generalization of association rules through disjunction. *Annals of Mathematics and Artificial Intelligence 59(2)*, 201–222.

He, Z. et X. Xu (2005). FP-OUTLIER : Frequent pattern based outlier detection. *Computer Science Information System 2(1)*, 103–118.

Jaccard, P. (1901). Étude comparative de la distribution orale dans une portion des Alpes et des Jura. *Bulletin de la Société Vaudoise des Sciences Naturelles 37*, 547–579.

Kim, W. Y., Y. K. Lee, et J. Han (2004). CCMINE : Efficient mining of confidence-closed correlated patterns. In *Proceedings of the 8th International Pacific-Asia Conference on Knowledge Data Discovery (PAKDD 2004), LNAI, volume 3056, Springer-Verlag, Sydney, Australie*, pp. 569–579.

Koh, Y. S. et N. Rountree (2010). *Rare Association Rule Mining and Knowledge Discovery : Technologies for Infrequent and Critical Event Detection*. IGI Global Publisher.

Lee, A. J. T., W. C. Lin, et C.-S. Wang (2006). Mining association rules with multi-dimensional constraints. *The Journal of Systems and Software 79(1)*, 79–92.

Lee, S. D. et L. De Raedt (2004). An efficient algorithm for mining string databases under constraints. In *Proceedings of the 3rd International Workshop on Knowledge Discovery in Inductive Databases (KDID 2004), LNCS, volume 3377, Springer-Verlag, Pisa, Italy*, pp. 108–129.

Lee, Y. K., W. Y. Kim, Y. D. Cai, et J. Han (2003). CoMine : efficient mining of correlated patterns. In *Proceedings of the 3rd International Conference on Data Mining (ICDM 2003), IEEE Computer Society Press, Melbourne, Florida, USA*, pp. 581–584.

Lei, J., P. Renqing, et P. Dingyu (2003). Tough constraint-based frequent closed itemsets mining. In *Proceedings of the 18th Symposium on Applied Computing (SAC 2003), ACM Press, New York, USA*, pp. 416–420.

Ma, S. et J. L. Hellerstein (2001). Mining mutually dependent patterns. In *Proceedings of the 1st International Conference on Data Mining (ICDM 2001), IEEE Computer Society Press, San Jose, California, USA*, pp. 409–406.

Mahmood, A. N., J. Hu, Z. Tari, et C. Leckie (2010). Critical infrastructure protection : Resource efficient sampling to improve detection of less frequent patterns in network traffic. *Journal of Network and Computer Applications 33(4)*, 491–502.

Mannila, H. et H. Toivonen (1997). Levelwise search and borders of theories in knowledge discovery. *Data Mining and Knowledge Discovery 3(1)*, 241–258.

Manning, A. M., D. J. Haglin, et J. A. Keane (2008). A recursive search algorithm for statistical disclosure assessment. *Data Mining and Knowledge Discovery 16(2)*, 165–196.

Okubo, Y., M. Haraguchi, et T. Nakajima (2010). Finding rare patterns with weak correlation constraint. In *Proceedings of the 10th IEEE International Conference on Data Mining (ICDM 2010) Workshop on Chance Discovery (IWCD 2010), Sydney, Australia, December 2010*, pp. 822–829.

Omiecinski, E. (2003). Alternative interest measures for mining associations in databases. *IEEE Transactions on Knowledge and Data Engineering 15*(1), 57–69.

Padmanabhan, B. et A. Tuzhilin (2006). On characterization and discovery of minimal unexpected patterns in rule discovery. *IEEE Transactions on Knowledge and Data Engineering 18(2)*, 202–216.

Pasquier, N., Y. Bastide, R. Taouil, G. Stumme, et L. Lakhal (2005). Generating a condensed representation for association rules. *Intelligent Information Systems 24(1)*, 25–60.

Pei, J. et J. Han (2004). Constrained frequent pattern mining : a pattern-growth view. *ACM-SIGKDD Explorations 4(1)*, 31–39.

Romero, C., J. R. Romero, J. M. Luna, et S. Ventura (2010). Mining rare association rules from e-learning data. In *Proceedings of the 3rd International Conference on Educational Data Mining (EDM 2010), Pittsburgh, PA, USA*, pp. 171–180.

Sandler, I. et A. Thomo (2010). Mining frequent highly-correlated item-pairs at very low support levels. In *Proceedings of the 10th SIAM International Conference on Data Mining (SDM 2010) Workshop on High Performance Analytics - Algorithms, Implementations, and Applications (PHPA 2010), Columbus, Ohio, USA*.

Segond, M. et C. Borgelt (2011). Item set mining based on cover similarity. In *Proceedings of the 15th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2011), LNCS, volume 6635, Springer-Verlag, Shenzhen, China*, pp. 493–505.

Surana, A., R. U. Kiran, et P. K. Reddy (2010). Selecting a right interestingness measure for rare association rules. In *Proceedings of the 16th International Conference on Management of Data (COMAD 2010), Nagpur, India*, pp. 115–124.

Szathmary, L., P. Valtchev, et A. Napoli (2010). Generating rare association rules using the minimal rare itemsets family. *International Journal of Software and Informatics 4(3)*, 219–238.

Taniar, D., W. Rahayu, V. Lee, et O. Daly (2008). Exception rules in association rule mining. *Applied Mathematics and Computation 205(2)*, 735–750.

Tanimoto, T. T. (1958). An elementary mathematical theory of classification and prediction. *Technical Report, I.B.M. Corporation Report*.

Weiss, G. M. (2004). Mining with rarity : A unifying framework. *ACM-SIGKDD Explorations 6(1)*, 7–19.

Xiong, H., P. N. Tan, et V. Kumar (2006). Hyperclique pattern discovery. *Data Mining and Knowledge Discovery. 13*(2), 219–242.

Summary

In the literature, works have been mainly dedicated to the extraction of frequent patterns. However, recently, rare pattern mining proves to be of added-value since these patterns allow conveying knowledge on rare and unexpected events. They are hence useful in several application fields. Nevertheless, a main claim related to rare pattern extraction is, on the one hand, their very high number and, on the other hand, the low quality of several mined patterns. The latter can indeed not present strong correlations between the items they contain. In order to overcome these limits, we propose to integrate the correlation measure *bond* aiming at only mining the set of rare patterns fulfilling this measure. A characterization of the resulting set, of rare correlated patterns, is then carried out based on the study of constraints of distinct types induced by the rarity and the correlation. In addition, based on the equivalence classes associated to a closure operator dedicated to the *bond* measure, we propose exact concise representations of rare correlated patterns.