Sémantiques et Calculs de Règles Descriptives dans une Relation n-aire

Kim-Ngan T. Nguyen, Loïc Cerf et Jean-François Boulicaut
Université de Lyon, CNRS, INRIA
INSA-Lyon, LIRIS Combining, UMR5205, F-69621, France

Résumé

La découverte de règles dans des données Booléennes a été particulièrement bien étudiée avec, notamment, le calcul de règles d'association pertinentes. Il s'agit alors de fouiller des relations binaires comme des relations *Transactions* × *Produits* ou plus généralement *Objets* × *Propriétés*. Dans cet article, nous étudions la généralisation de ce concept de règle d'association au contexte des relations d'arité arbitraire. En effet, de nombreux jeux de données sont des représentations de relations n-aires (avec un nombre de dimensions $n > 2$). Cependant, peu de propositions existent pour adosser des processus de découverte de connaissances sur des règles descriptives dont les prémisses et les conclusions peuvent impliquer des sous-ensembles de n'importe lesquelles des n dimensions. L'un des verrous consiste à fixer la sémantique de telles règles via des mesures d'intérêt objectives qu'il a fallu concevoir (e.g., les notions de confiance naturelle ou exclusive). Nous proposons un premier algorithme d'extraction ayant été validé sur des données réelles.

Mots clés : Règles descriptives, Données multidimensionnelles, Tenseur Booléen.

1 Introduction

La fouille de relations binaires a mobilisé énormément de chercheurs et de ressources. Il s'agit, par exemple, de relations *Transactions* × *Produits* (on parle aussi de données transactionnelles) ou plus généralement de relations *Objets* × *Propriétés* où les deux dimensions peuvent être de grandes tailles. De nombreuses propositions permettent aujourd'hui d'alimenter des processus de découverte de connaissances à partir de telles données. Nous nous intéressons aux méthodes descriptives basées sur des calculs de régularités ou de motifs locaux. Il peut s'agir d'ensembles fréquents (voir, e.g., [2, 16]), d'ensembles fermés ou de concepts formels (voir, e.g., [9, 20]), de règles d'association (voir, e.g., [2, 3]) ou encore de leurs généralisations comme, par exemple, l'introduction de négations [16, 18] ou la découverte de règles dans un contexte multi-relationnel [8, 12]. Il existe aujourd'hui un savoir-faire algorithmique pour calculer efficacement de nombreux types de motifs dans de grandes relations binaires. Ceci étant, de nombreux jeux de données se présentent naturellement comme des relations n-aires avec, par exemple,

l'ajout de dimensions spatiales ou temporelles sur des relations *Transactions* × *Produits* qui deviennent des relations *Transactions* × *Produits* × *Date* × *Lieu de vente*.

Étendre les méthodes de fouille de relations binaires au contexte des relation d'arité arbitraire paraît donc être une direction de recherche importante et encore peu étudiée. Le problème est que l'extension aux relations n-aires est plus ou moins difficile et que nous devons considérer trois problèmes majeurs dans la fouille de données non supervisée au moyen de motifs (ou des règles descriptives qui peuvent en être dérivées).

1. Quelle est la sémantique du domaine de motif ? Autrement dit, quelles sont les formes que peuvent prendre les motifs dans des relations n-aires et quels sont les mesures qui vont permettre d'en déterminer l'intérêt a priori ? Si l'on veut spécifier ce qu'est une règle d'association [2] dans ce contexte des relations n-aires, que deviennent les classiques mesures de fréquence et de confiance ?
2. Quels sont les mécanismes qui vont permettre de spécifier les attentes de l'analyste et donc l'intérêt subjectif ? Depuis quelques années, de nombreux chercheurs développent le cadre de la fouille de données sous contraintes pour lequel des combinaisons Booléennes de contraintes primitives peuvent spécifier déclarativement des propriétés souhaitées sur les motifs solutions (voir, e.g., [5]). Il faudrait donc idéalement identifier les "bonnes" contraintes primitives.
3. Quels sont les moyens de calculs qui vont permettre de calculer les motifs solutions c'est-à-dire satisfaisant les contraintes posées ? Si possible, on souhaite réaliser des calculs corrects et complets qui délivrent tous les motifs solutions et seulement ceux-là. Il faut pouvoir passer à l'échelle au regard du nombre de dimensions et de la taille (nombre de valeurs) de chacune d'entre elles.

Ainsi, étendre la sémantique des motifs ensemblistes comme des concepts formels (couples d'ensembles fermés sur chacune des deux dimensions) au contexte des relations n-aires est trivial d'un point de vue déclaratif (spécification a priori des critères d'intérêts objectifs et subjectifs) mais difficile sur un plan calculatoire [13, 11, 6]. Par contre, et c'est l'objet de cet article, définir la sémantique de ce que pourraient être des règles d'association dans des relations n-aires s'est révélé délicat. En fait, depuis la proposition initiale de cette tâche prototypique en fouille de données [2], la sémantique des règles d'association a été assez peu étudiée et formalisée (avec quelques exceptions comme [1]). Bien qu'il s'agisse d'un type de motif simple, on note que des notions importantes pour la sémantique des règles (e.g., les concepts de fréquence ou de contre-exemples) peuvent connaître des définitions différentes selon les auteurs.

Lorsque l'on travaille sur des relations n-aires, il va falloir redéfinir et le langage des motifs et ce que peuvent être de telles mesures lorsque les prémisses et les conclusions des règles peuvent porter sur des sous-ensembles de n'importe lesquelles des dimensions. Ainsi, notre contribution principale consiste à concevoir la sémantique des règles via des mesures d'intérêt comme les notions de confiance naturelle ou de confiance exclusive. Ce travail sur la sémantique des règles fait l'objet d'une étude théorique mais aussi d'une validation empirique au moyen d'un cas réel et donc d'exemples de motifs découverts effectivement pertinents. Nous considérons des extractions sur un jeu de données correspondant à une relation ternaire sur les consultations de distributions GNU/Linux.

Notre seconde contribution est algorithmique et concerne la conception d'un premier algorithme d'extraction efficace pour calculer les règles a priori intéressantes. Il s'appuie

	p_1	p_2	p_3	p_4	p_1	p_2	p_3	p_4	p_1	p_2	p_3	p_4	p_1	p_2	p_3	p_4
c_1	**1**	1	1		1	1	1		1		1		1	1		1
c_2	1	1		1	1	1				1	1	1			1	1
c_3	1	1			1					1	1	1			1	1
c_4		1		1			1	1			1		1	1	1	
c_5												1				1
	s_1				s_2				s_3				s_4			

TAB. 1: $\mathcal{R}_E \subseteq \{p_1, p_2, p_3, p_4\} \times \{s_1, s_2, s_3, s_4\} \times \{c_1, c_2, c_3, c_4, c_5\}$

sur les principes qui viennent d'être proposés pour le calcul de motifs multidimensionnels fermés [6, 7]. Nous décrivons l'algorithme et nous établissons quelques unes de ses propriétés. Son comportement expérimental est également étudié.

Dans la Section 2, nous introduisons les notions de base qui vont permettre la construction du domaine de motif des règles dans la Section 3. Nous décrivons notre algorithme de calcul des règles a priori intéressantes dans la Section 4. La Section 5 est dédiée aux études empiriques. La Section 6 discute des travaux connexes. Enfin, la Section 7 est une brève conclusion.

2 Définitions préliminaires

Soit n ensembles finis supposés disjoints (sans perte de généralité) $\{D^1, \ldots, D^n\} = \mathcal{D}$. Nous notons $\mathcal{R} \subseteq D^1 \times \cdots \times D^n$ la relation n-aire à partir de laquelle on souhaite découvrir des associations. Considérons un exemple jouet de relation ternaire, $\mathcal{R}_E$, représentée dans Tab. 1. $\mathcal{R}_E$ relie des produits de $D^1 = \{p_1, p_2, p_3, p_4\}$ achetés au cours des saisons de $D^2 = \{s_1, s_2, s_3, s_4\}$ par des clients de $D^3 = \{c_1, c_2, c_3, c_4, c_5\}$. Chaque '1' dans Tab. 1 se trouve à l'intersection de trois éléments $(p_i, s_j, c_k) \in D^1 \times D^2 \times D^3$ formant un triplet présent dans $\mathcal{R}_E$. Ainsi le produit p_1 est acheté à la saison s_1 par le client c_1 (le '1' correspondant est en gras dans Tab. 1) alors que le client c_4 ne se procure ce produit que lors de la saison s_4.

Les associations ou co-occurrences qui nous intéressent ne concernent que certains domaines $\mathcal{D}' \subseteq \mathcal{D}$ de la relation. Par exemple, étant donné $\mathcal{R}_E$, on peut s'intéresser aux associations impliquant les produits et les saisons ($\mathcal{D}' = \{D^1, D^2\}$). Sans perte de généralité, nous supposons que $\mathcal{D}' = \{D^1, \ldots, D^{|\mathcal{D}'|}\}$. Nous appelons *association* sur $\mathcal{D}'$ le produit cartésien de sous-ensembles non vides des domaines de $\mathcal{D}'$.

Définition 1 (Association) *$\forall \mathcal{D}' = \{D^1, \ldots, D^{|\mathcal{D}'|}\} \subseteq \mathcal{D}$, $X^1 \times \cdots \times X^{|\mathcal{D}'|}$ est une association sur $\mathcal{D}'$ si et seulement si $\forall i = 1..|\mathcal{D}'|, X^i \subseteq D^i \wedge X^i \neq \emptyset$.*

Par convention, si $\mathcal{D}'$ est vide, la seule *association* sur $\mathcal{D}'$ est l'ensemble vide noté $\emptyset$. $\times_{D^i \in \mathcal{D} \setminus \mathcal{D}'} D^i$, est appelé *domaine support*. Par exemple, dans $\mathcal{R}_E$, D^3 est le domaine support d'une association sur $\{D^1, D^2\}$. Le domaine support permet de définir le support d'une association. Cette définition utilise l'opérateur de concaténation noté $\cdot$. On a, par exemple, $(p_2, s_1) \cdot (c_2) = (p_2, s_1, c_2)$.

Définition 2 (Support d'une association) *$\forall \mathcal{D}' \subseteq \mathcal{D}$, soit X une association sur $\mathcal{D}'$, son support noté $s(X)$ est :*

$$s(X) = \{t \in \times_{D^i \in \mathcal{D} \setminus \mathcal{D}'} D^i \mid \forall x \in X, x \cdot t \in \mathcal{R}\} \ .$$

On peut noter qu'une association impliquant tous les n domaines ($\mathcal{D}' = \mathcal{D}$) est soit vraie (tous les n-uplets qu'elle contient appartiennent à $\mathcal{R}$), soit fausse (au moins un des n-uplets qu'elle contient n'appartient *pas* à $\mathcal{R}$). Nous n'avons donc pas de graduation possible de sa qualité. Dans ce cas particulier, en utilisant la convention $\times_{D^i \in \emptyset} D^i = \{\epsilon\}$ (où ϵ est le mot vide), les associations possibles ont bien, respectivement, soit un support d'un élément soit un support vide. Un second cas extrême, et peu intéressant, correspond à $s(\emptyset) = \mathcal{R}$. Le support d'une association généralise celui d'un *itemset* dans une relation binaire (cas où $n = 2$ et $\mathcal{D}' = \{D^1\}$).

Dans la suite, la notion de support d'une association est très utilisée. Donnons quelques définitions complémentaires pour exprimer la sémantique d'une règle d'association dans une relation n-aire.

Définition 3 (Composante) *$\forall \mathcal{D}' = \{D^1, \ldots, D^{|\mathcal{D}'|}\} \subseteq \mathcal{D}$, soit $X = X^1 \times \cdots \times X^{|\mathcal{D}'|}$ une association sur $\mathcal{D}'$. $\forall D^i \in \mathcal{D}$, la composante de X sur D^i, notée $\pi_{D^i}(X)$, est X^i si $D^i \in \mathcal{D}'$, $\emptyset$ sinon.*

Définition 4 (Union d'associations) *$\forall \mathcal{D}_X \subseteq \mathcal{D}$ et $\forall \mathcal{D}_Y \subseteq \mathcal{D}$, soit X une association sur $\mathcal{D}_X$ et Y une association sur $\mathcal{D}_Y$, l'union de X et Y notée $X \sqcup Y$ est l'association sur $\mathcal{D}_X \cup \mathcal{D}_Y$ pour laquelle $\forall D^i \in \mathcal{D}, \pi_{D^i}(X \sqcup Y) = \pi_{D^i}(X) \cup \pi_{D^i}(Y)$.*

Définition 5 (Complément d'associations) *$\forall \mathcal{D}_X \subseteq \mathcal{D}$ et $\forall \mathcal{D}_Y \subseteq \mathcal{D}$, soit X une association sur $\mathcal{D}_X$ et Y une association sur $\mathcal{D}_Y$, le complément de X dans Y noté $Y \setminus X$ est l'association sur $\{D^i \in \mathcal{D}_Y \mid \pi_{D^i}(Y) \not\subseteq \pi_{D^i}(X)\}$ telle que $\forall D^i \in \mathcal{D}, \pi_{D^i}(Y \setminus X) = \pi_{D^i}(Y) \setminus \pi_{D^i}(X)$.*

Définition 6 (Inclusion d'associations) *$\forall \mathcal{D}_X \subseteq \mathcal{D}$ et $\forall \mathcal{D}_Y \subseteq \mathcal{D}$, soit X une association sur $\mathcal{D}_X$ et Y une association sur $\mathcal{D}_Y$, l'inclusion des associations est notée $X \sqsubseteq Y$. On a $X \sqsubseteq Y \Leftrightarrow \forall D^i \in \mathcal{D}, \pi_{D^i}(X) \subseteq \pi_{D^i}(Y)$.*

L'anti-monotonie du support est préservée dans le cadre plus général des associations et en utilisant la notion d'inclusion que nous venons de définir.

Théorème 1 (Anti-monotonie du support) *$\forall \mathcal{D}_X \subseteq \mathcal{D}$ et $\forall \mathcal{D}_Y \subseteq \mathcal{D}$, soit X une association sur $\mathcal{D}_X$ et Y une association sur $\mathcal{D}_Y$, on a :*

$$X \sqsubseteq Y \Rightarrow |s(X)| \geq |s(Y)| \ .$$

Preuve en annexe.

Dans $\mathcal{R}_E$, représentée Tab. 1, $\{p_1, p_2\} \times \{s_1\}$ et $\{p_1, p_2\} \times \{s_1, s_2\}$ sont deux associations sur $\{D^1, D^2\}$. Par contre, $\{p_1, p_2\}$ n'est pas une association sur $\{D^1, D^2\}$ car sa projection sur D^2 est vide. C'est une association sur $\{D^1\}$.

- $s(\{p_1, p_2\} \times \{s_1\}) = \{c_1, c_2, c_3\}$;
- $s(\{p_1, p_2\} \times \{s_1, s_2\}) = \{c_1, c_2\}$;
- $s(\{p_1, p_2\}) = \{(s_1, c_1), (s_1, c_2), (s_1, c_3), (s_2, c_1), (s_2, c_2), (s_4, c_1), (s_4, c_4)\}$.

Par ailleurs, comme $\{p_1, p_2\} \sqsubseteq \{p_1, p_2\} \times \{s_1\} \sqsubseteq \{p_1, p_2\} \times \{s_1, s_2\}$, on observe bien $|s(\{p_1, p_2\})| \geq |s(\{p_1, p_2\} \times \{s_1\})| \geq |s(\{p_1, p_2\} \times \{s_1, s_2\})|$.

3 Règle d'association dans une relation n-aire

3.1 Définition

Étant donné une relation n-aire sur $\mathcal{D}$ et $\mathcal{D}' \subseteq \mathcal{D}$ l'ensemble des domaines choisis par l'analyste, une règle d'association sur $\mathcal{D}'$ est un couple d'associations sur des ensembles de domaines qui peuvent être différents mais dont l'union doit être $\mathcal{D}'$. Le domaine support de la règle est $\times_{D^i \in \mathcal{D} \setminus \mathcal{D}'} D^i$.

Définition 7 (Règle d'association multi-dimensionnelle) *$\forall \mathcal{D}' \subseteq \mathcal{D}$, $X \to Y$ est une règle d'association sur $\mathcal{D}'$ si et seulement si $\exists \mathcal{D}_X \subseteq \mathcal{D}$ et $\exists \mathcal{D}_Y \subseteq \mathcal{D}$ tel que X est une association sur $\mathcal{D}_X$, Y est une association sur $\mathcal{D}_Y$, et $\mathcal{D}_X \cup \mathcal{D}_Y = \mathcal{D}'$.*

Dans $\mathcal{R}_E$, $\{p_1, p_2\} \to \{s_1, s_2\}$ et $\{p_4\} \times \{s_3, s_4\} \to \{p_3\}$ sont deux exemples de règles d'association sur $\{D^1, D^2\}$. La règle $\{p_1\} \to \{p_2\}$ n'est pas une règle d'association sur $\{D^1, D^2\}$ car aucun élément de D^2 n'apparaît dans le membre gauche (la prémisse) ou le membre droit (la conclusion). Par contre, $\{p_1\} \to \{p_2\}$ est bien une règle d'association sur $\{D^1\}$.

Dans le cas binaire, la sémantique classique d'une règle d'association repose sur les mesures de fréquence et de confiance et l'intérêt a priori d'une règle est spécifié au moyen de seuils : une règle a priori intéressante satisfait une conjonction de contraintes spécifiant que et sa fréquence et sa confiance doivent être supérieures à des seuils fournis par les analystes [2]. Une règle est fréquente si elle se vérifie sur un grand nombre d'éléments du domaine support. Plus précisément, l'union de la prémisse et de la conclusion de la règle a pour support un ensemble contenant un nombre suffisant d'éléments. Une règle est valide au sens d'une confiance suffisante si la probabilité conditionnelle d'observer la conclusion lorsque l'on observe la prémisse est suffisamment grande. En fait, dans le contexte des règles multi-dimensionnelles, une seule définition de la fréquence d'une règle va s'imposer naturellement. Par contre, il va être difficile de définir la confiance d'une règle dans le cas où l'association en conclusion est définie sur un ensemble de domaines qui n'est pas inclus dans celui de la prémisse.

3.2 Définition de la fréquence

La fréquence (relative) d'une règle d'association est, dans le domaine support, la proportion d'éléments dans le support de l'union de la prémisse et de la conclusion.

Définition 8 (Fréquence d'une règle) *$\forall \mathcal{D}' \subseteq \mathcal{D}$, soit $X \to Y$ une règle d'association sur $\mathcal{D}'$. Sa fréquence, notée $f(X \to Y)$, est :*

$$f(X \to Y) = \frac{|s(X \sqcup Y)|}{|\times_{D^i \in \mathcal{D} \setminus \mathcal{D}'} D^i|} \ .$$

Dans $\mathcal{R}_E$, en appliquant en cascade les Définitions 8, 4 et 2, nous obtenons :
- $f(\{p_1, p_2\} \to \{s_1, s_2\}) = \frac{|s(\{p_1,p_2\} \times \{s_1,s_2\})|}{|D^3|} = \frac{|\{c_1,c_2\}|}{|\{c_1,c_2,c_3,c_4,c_5\}|} = \frac{2}{5}$;
- $f(\{p_4\} \times \{s_3, s_4\} \to \{p_3\}) = \frac{|s(\{p_3,p_4\} \times \{s_3,s_4\})|}{|D^3|} = \frac{|\{c_2,c_3\}|}{|\{c_1,c_2,c_3,c_4,c_5\}|} = \frac{2}{5}$.

3.3 Définition de la confiance

3.3.1 Difficulté à définir la confiance

Est-il possible de généraliser facilement le concept de confiance d'une règle d'association dans une relation binaire à notre nouveau contexte, et ainsi de vouloir attribuer à une règle $X \rightarrow Y$ la mesure de confiance $\frac{|s(X \sqcup Y)|}{|s(X)|}$? Lorsque X et $X \sqcup Y$ sont des associations sur le même ensemble de domaines (leurs domaines support sont donc les mêmes), cette définition est souhaitable. Elle est une proportion d'éléments d'un même domaine support. Ainsi, dans $\mathcal{R}_E$, cette définition de la confiance devrait attribuer à la règle $\{p_4\} \times \{s_3, s_4\} \rightarrow \{p_3\}$ la valeur $\frac{|s(\{p_3,p_4\} \times \{s_3,s_4\})|}{|s(\{p_4\} \times \{s_3,s_4\})|} = \frac{|\{c_2,c_3\}|}{|\{c_2,c_3,c_5\}|} = \frac{2}{3}$, ce qui correspond à une proportion de *clients*. Cela signifie que, parmi ceux qui achètent le produit p_4 à la fois aux saisons s_3 et s_4, la plupart achète aussi le produit p_3 durant ces saisons.

Cependant, cette sémantique n'est pas satisfaisante pour une règle où l'association en conclusion est définie sur un ensemble de domaines qui n'est pas inclus dans celui de l'association en prémisse. En effet, $s(X \sqcup Y)$ et $s(X)$ sont alors des ensembles disjoints et mettre leurs cardinaux en rapport n'a aucun sens. Par exemple, considérons la règle $\{p_1, p_2\} \rightarrow \{s_1, s_2\}$ dans $\mathcal{R}_E$. On a $s(\{p_1, p_2\} \times \{s_1, s_2\}) = \{c_1, c_2\}$, qui est un ensemble de *clients*, et $s(\{p_1, p_2\}) = \{(s_1, c_1), (s_1, c_2), (s_1, c_3), (s_2, c_1), (s_2, c_2), (s_4, c_1), (s_4, c_4)\}$, qui est un ensemble de couples $(saison, client)$.

Nous avons donc travaillé à la définition de mesures de confiance fondées d'un point de vue sémantique. Les définitions que nous proposons correspondent chacune à $\frac{|s(X \sqcup Y)|}{|s(X)|}$ lorsque l'association conclusion est définie sur un ensemble de domaines inclus dans celui de l'association prémisse. En particulier, elles sont des généralisations de la mesure de confiance introduite dans [2].

3.3.2 Confiance exclusive

Calculer la confiance de $X \rightarrow Y$ pose donc un problème lorsque X est définie sur un ensemble $\mathcal{D}_X$ inclus *strictement* dans celui $\mathcal{D}'$ de $X \sqcup Y$. Un facteur correctif permet néanmoins de rendre $|s(X)|$ et $|s(X \sqcup Y)|$ comparables. Il s'agit de multiplier $|s(X \sqcup Y)|$ par les cardinaux de ses composantes sur les domaines absents de $\mathcal{D}_X$.

Définition 9 (Confiance exclusive) *$\forall \mathcal{D}' \subseteq \mathcal{D}$, soit $X \rightarrow Y$ une règle d'association sur $\mathcal{D}'$ et notons $\mathcal{D}_X$ l'ensemble de domaines sur lequel X est défini, sa confiance exclusive notée $c_{exclusive}(X \rightarrow Y)$ est :*

$$c_{exclusive}(X \rightarrow Y) = \frac{|s(X \sqcup Y)| \times | \times_{D^i \in \mathcal{D}' \setminus \mathcal{D}_X} \pi_{D^i}(Y)|}{|s(X)|} \ .$$

Lorsque X est une association sur $\mathcal{D}'$, la confiance exclusive de $X \rightarrow Y$ vaut $\frac{|s(X \sqcup Y)|}{|s(X)|}$ sous la convention $\times_{D^i \in \emptyset} \pi_{D^i}(Y) = \{\epsilon\}$. Le facteur correctif, $| \times_{D^i \in \mathcal{D}' \setminus \mathcal{D}_X} \pi_{D^i}(Y)|$, appliqué à $|s(X \sqcup Y)|$ permet de comptabiliser les éléments de $s(X \sqcup Y)$ « de la même façon au numérateur et au dénominateur de la fraction ».

Par exemple considérons la règle $\{p_1, p_2\} \rightarrow \{s_1, s_2\}$ dans $\mathcal{R}_E$ et supposons que l'achat d'un client en une saison s'appelle une transaction. On trouve qu'il n'y a que deux clients $\{c_1, c_2\}$ qui achètent les deux produits p_1 et p_2 à la fois aux saisons s_1 et s_2. Dans

ce cas, la somme des transactions pour lesquelles les produits p_1 et p_2 sont achetés par les clients c_1 et c_2 au moment des saisons s_1 et s_2 est $|\{c_1, c_2\}| \times |\{s_1, s_2\}| = 4$. La somme des transactions pour lesquelles les produits p_1 et p_2 sont achetés en n'importe quelle saison est $|\{(s_1, c_1), (s_1, c_2), (s_1, c_3), (s_2, c_1), (s_2, c_2), (s_4, c_1), (s_4, c_4)\}| = 7$. La confiance exclusive de la règle vaut donc :

$$\begin{aligned} c_{\text{exclusive}}(\{p_1, p_2\} \rightarrow \{s_1, s_2\}) &= \frac{|s(\{p_1,p_2\}\times\{s_1,s_2\})|\times|\{s_1,s_2\}|}{|s(\{p_1,p_2\})|} \\ &= \frac{|\{c_1,c_2\}|\times|\{s_1,s_2\}|}{|\{(s_1,c_1),(s_1,c_2),(s_1,c_3),(s_2,c_1),(s_2,c_2),(s_4,c_1),(s_4,c_4)\}|} \\ &= \frac{4}{7} \ . \end{aligned}$$

Le fait que le client c_3 achète les deux produits p_1 et p_2 à la saison s_1 mais qu'il ne les achète pas ensemble à la saison s_2 fait aussi « baisser » la confiance en ce que les clients aiment bien acheter ces produits à la fois aux saisons s_1 et s_2. Le fait que le client c_1 achète ces produits en saison s_4 fait « baisser » la confiance sur le fait que l'on n'aime les acheter qu'aux saisons s_1 et s_2. Si cette confiance valait 1 et donc la valeur maximale, cela voudrait dire que les clients appréciant les deux produits p_1 et p_2 achètent ces produits aux saisons s_1 et s_2 mais aussi qu'ils ne les achètent pas pendant les autres saisons. C'est pourquoi nous parlons de *confiance exclusive*.

Cette exclusivité présente un défaut dommageable à une extraction efficace des règles d'association valides, c'est-à-dire présentant une confiance supérieure à un seuil fixé par l'analyste : $X \mapsto c_{\text{exclusive}}(X \rightarrow Y \setminus X)$ n'est pas une fonction croissante sur un ensemble $X \sqsubseteq Y$ ordonné par $\sqsubseteq$. Par exemple, considérons les règles $\{s_3\} \rightarrow \{p_2, p_3, p_4\}$ et $\{s_3\} \times \{p_3\} \rightarrow \{p_2, p_4\}$ dans $\mathcal{R}_E$. On a bien $\{s_3\} \sqsubseteq \{s_3\} \times \{p_3\}$. Pourtant $c_{\text{exclusive}}(\{s_3\} \rightarrow \{p_2, p_3, p_4\}) = \frac{6}{10}$ et $c_{\text{exclusive}}(\{s_3\} \times \{p_3\} \rightarrow \{p_2, p_4\}) = \frac{2}{4}$. Donc $c_{\text{exclusive}}(\{s_3\} \times \{p_3\} \rightarrow \{p_2, p_4\}) < c_{\text{exclusive}}(\{s_3\} \rightarrow \{p_2, p_3, p_4\})$. Cette absence de propriété n'ôte pas la pertinence de la confiance exclusive. Elle pénalise une règle d'association dont les éléments du domaine support la vérifiant permettent, individuellement, de conclure sur d'autres éléments que ceux en conclusion de la règle. Par conséquent, elle favorise la découverte d'une règle d'association concluant sur un maximum d'éléments. La difficulté algorithmique à extraire des règles d'association sous contrainte de confiance exclusive minimale et la volonté d'une mesure plus facile à interpréter nous a conduit à construire une nouvelle généralisation de la mesure de confiance.

3.3.3 Confiance naturelle

Rappelons que la définition de la confiance de $X \rightarrow Y$ est problématique lorsque le domaine support servant au calcul du support de X, $\times_{D^i \in \mathcal{D} \setminus \mathcal{D}_X} D^i$, est différent du domaine support $\times_{D^i \in \mathcal{D} \setminus \mathcal{D}'} D^i$ utilisé pour obtenir $s(X \sqcup Y)$. La confiance dite naturelle repose sur l'idée de ramener le support de X à un sous-ensemble de $\times_{D^i \in \mathcal{D} \setminus \mathcal{D}'} D^i$. La confiance de $X \rightarrow Y$ est alors une proportion d'éléments de $\times_{D^i \in \mathcal{D} \setminus \mathcal{D}'} D^i$ et se voit qualifiée de *naturelle*. Le prix à payer est la nécessité d'une nouvelle définition du support spécifique aux prémisses des règles et dépendant aussi de leurs conclusions.

Définition 10 (Support naturel d'une prémisse) *$\forall \mathcal{D}' \subseteq \mathcal{D}$, soit $X \rightarrow Y$ une règle d'association sur $\mathcal{D}'$, le support naturel de X noté $s_{\mathcal{D} \setminus \mathcal{D}'}(X)$ est :*

$$s_{\mathcal{D} \setminus \mathcal{D}'}(X) = \{t \in \times_{D^i \in \mathcal{D} \setminus \mathcal{D}'} D^i \mid \exists u \in \times_{D^i \in \mathcal{D}' \setminus \mathcal{D}_X} D^i \text{ tel que } \forall x \in X, x \cdot u \cdot t \in \mathcal{R}\} \ .$$

où $\mathcal{D}_X$ est l'ensemble des domaines de définition de X et $x \cdot u \cdot t$ est la concaténation de x, u et t (quitte à changer l'indexation des domaines de sorte que ceux dans $\mathcal{D}_X$ soient les premiers).

Définition 11 (Confiance naturelle) *$\forall \mathcal{D}' \subseteq \mathcal{D}$, soit $X \to Y$ une règle d'association sur $\mathcal{D}'$, sa confiance naturelle notée $c_{naturelle}(X \to Y)$ est :*

$$c_{naturelle}(X \to Y) = \frac{|s(X \sqcup Y)|}{|s_{\mathcal{D} \setminus \mathcal{D}'}(X)|} \ .$$

Lorsque X est une association sur $\mathcal{D}'$, comme pour la confiance exclusive, la confiance naturelle de $X \to Y$ vaut $\frac{|s(X \sqcup Y)|}{|s(X)|}$ sous la convention $\times_{D^i \in \emptyset} D^i = \{\epsilon\}$ que nous avons déjà utilisée.

Dans $\mathcal{R}_E$, considérons à nouveau la règle $\{p_1, p_2\} \to \{s_1, s_2\}$. Les clients qui achètent les produits p_1 et p_2 ensemble (lors d'au moins une saison) sont $\{c_1, c_2, c_3, c_4\}$. Ceux qui les achètent ensemble à la fois en s_1 et en s_2 sont $\{c_1, c_2\}$. La confiance naturelle la règle vaut donc :

$$c_{\text{naturelle}}(\{p_1, p_2\} \to \{s_1, s_2\}) = \frac{|s(\{p_1, p_2\} \times \{s_1, s_2\})|}{|s_{\{D^3\}}(\{p_1, p_2\})|} = \frac{|\{c_1, c_2\}|}{|\{c_1, c_2, c_3, c_4\}|} = \frac{2}{4} \ .$$

La confiance naturelle mesure ainsi, parmi les clients ayant au moins une fois acheté p_1 et p_2 ensemble, la proportion qui les achète ensemble à la fois aux deux saisons s_1 et s_2. À la différence de la confiance exclusive, les clients vérifiant la règle pourraient, par ailleurs, acheter p_1 et p_2 au cours d'autres saisons sans que cela ne fasse « baisser » la confiance naturelle. De plus, la confiance naturelle a une bonne propriété lui permettant, contrairement à la confiance exclusive, un élagage de l'espace de recherche lors du calcul complet des règles à forte confiance :

Théorème 2 (Croissance de $X \mapsto c_{\textbf{naturelle}}(X \to Y \setminus X)$ selon $\sqsubseteq$) *Soit $X \to Y \setminus X$ et $X' \to Y \setminus X'$ deux règles d'association sur $\mathcal{D}'$, on a :*

$$X \sqsubseteq X' \sqsubseteq Y \Rightarrow c_{naturelle}(X \to Y \setminus X) \leq c_{naturelle}(X' \to Y \setminus X') \ .$$

Preuve en annexe.

Dans $\mathcal{R}_E$, $\{p_1, p_2\} \to \{s_1, s_2\}$ et $\{p_1, p_2\} \times \{s_1\} \to \{s_2\}$ sont deux exemples de règles d'association sur $\{D^1, D^2\}$. Leurs confiances naturelles sont :

- $c_{\text{naturelle}}(\{p_1, p_2\} \to \{s_1, s_2\}) = \frac{|s(\{p_1,p_2\} \times \{s_1,s_2\})|}{|s_{\{D^3\}}(\{p_1,p_2\})|} = \frac{|\{c_1,c_2\}|}{|\{c_1,c_2,c_3,c_4\}|} = \frac{2}{4}$;
- $c_{\text{naturelle}}(\{p_1, p_2\} \times \{s_1\} \to \{s_2\}) = \frac{|s(\{p_1,p_2\} \times \{s_1,s_2\})|}{|s_{\{D^3\}}(\{p_1,p_2\} \times \{s_1\})|} = \frac{|\{c_1,c_2\}|}{|\{c_1,c_2,c_3\}|} = \frac{2}{3}$.

Ces deux règles illustrent le Théorème 2. En effet, on a $\{p_1, p_2\} \sqsubseteq \{p_1, p_2\} \times \{s_1\} \sqsubseteq \{p_1, p_2\} \times \{s_1, s_2\}$ et $c_{\text{naturelle}}(\{p_1, p_2\} \to \{s_1, s_2\}) \leq c_{\text{naturelle}}(\{p_1, p_2\} \times \{s_1\} \to \{s_2\})$. Dans la Section 4.3, nous utilisons ce théorème pour élaguer des sous-espaces de recherche où nous sommes certain qu'aucune règle ne pourra satisfaire une contrainte de confiance naturelle minimale.

3.4 Propriétés

Dans le cas particulier $n = 2$, la sémantique des règles d'association à confiance 1 (et sans contrainte de fréquence) est *bien formée*. Par « bien formée », il faut entendre que les axiomes décrits dans [4] sont vérifiés et que l'on peut donc raisonner sur des collections de règles. Nous allons voir que la vérification de deux de ces axiomes n'est pas souhaitable dans le contexte des relations d'arité arbitraire. En effet, ils sont violés par des règles n'impliquant pas, dans leurs conclusions, d'autres dimensions que celles en prémisses, c'est à dire par des règles pour lesquelles la définition de la confiance généralise naturellement la mesure analogue de qualité des règles d'association dans des relations binaires.

Axiome 1 (Réflexivité) $\forall \mathcal{D}_X \subseteq \mathcal{D}$, $\forall \mathcal{D}_Y \subseteq \mathcal{D}$, *pour toutes les associations* X *sur* $\mathcal{D}_X$ *et* Y *sur* $\mathcal{D}_Y$ *on a :*

$$Y \sqsubseteq X \Rightarrow c(X \rightarrow Y) = 1 \ .$$

Axiome 2 (Augmentation) $\forall \mathcal{D}' \subseteq \mathcal{D}$, $\forall \mathcal{D}_Z \subseteq \mathcal{D}'$, *pour toute règle d'association* $X \rightarrow Y$ *sur* $\mathcal{D}'$ *et toute association* Z *sur* $\mathcal{D}_Z$, *on a :*

$$c(X \rightarrow Y) = 1 \Rightarrow c(X \sqcup Z \rightarrow Y \sqcup Z) = 1 \ .$$

Axiome 3 (Transitivité) $\forall \mathcal{D}' \subseteq \mathcal{D}$, *pour toutes les règles d'association* $X \rightarrow Y$ *et* $Y \rightarrow Z$ *sur* $\mathcal{D}'$, *on a :*

$$c(X \rightarrow Y) = 1 \wedge c(Y \rightarrow Z) = 1 \Rightarrow c(X \rightarrow Z) = 1 \ .$$

Que ce soit avec la sémantique exclusive ou avec la sémantique naturelle, on démontre directement que l'axiome de réflexivité est vérifié. En effet, si $Y \sqsubseteq X$, alors $X \sqcup Y = X$ et $s_{\mathcal{D}_X}(X) = s(X)$ d'où $c_{\text{exclusive}}(X \rightarrow Y) = c_{\text{naturelle}}(X \rightarrow Y) = \frac{|s(X \sqcup Y)|}{|s(X)|} = \frac{|s(X)|}{|s(X)|} = 1$.

En revanche, les axiomes d'augmentation et de transitivité ne sont plus justes. Par exemple, dans $\mathcal{R}_E$, $\{p_1\} \times \{s_1\} \rightarrow \{s_2\}$ est une règle sur $\{D^1, D^2\}$, $c_{\text{exclusive}}(\{p_1\} \times \{s_1\} \rightarrow \{s_2\}) = c_{\text{naturelle}}(\{p_1\} \times \{s_1\} \rightarrow \{s_2\}) = \frac{|\{c_1, c_2, c_3\}|}{|\{c_1, c_2, c_3\}|} = 1$. Pourtant, cette règle augmentée de $\{p_2\}$, i.e., $\{p_1, p_2\} \times \{s_1\} \rightarrow \{p_2\} \times \{s_2\}$, a pour confiance $c_{\text{exclusive}}(\{p_1, p_2\} \times \{s_1\} \rightarrow \{p_2\} \times \{s_2\}) = c_{\text{naturelle}}(\{p_1, p_2\} \times \{s_1\} \rightarrow \{p_2\} \times \{s_2\}) = \frac{|\{c_1, c_2\}|}{|\{c_1, c_2, c_3\}|} = \frac{2}{3}$.

Pour discuter l'axiome de transitivité, considérons les règles $\{p_1\} \times \{s_3\} \rightarrow \{p_3\} \times \{s_1, s_2\}$ et $\{p_3\} \times \{s_1, s_2\} \rightarrow \{p_2\}$ dans $\mathcal{R}_E$. Il s'agit de deux règles d'association sur $\{D^1, D^2\}$. On a $c_{\text{exclusive}}(\{p_1\} \times \{s_3\} \rightarrow \{p_3\} \times \{s_1, s_2\}) = c_{\text{naturelle}}(\{p_1\} \times \{s_3\} \rightarrow \{p_3\} \times \{s_1, s_2\}) = \frac{|\{c_1\}|}{|\{c_1\}|} = 1$ et $c_{\text{exclusive}}(\{p_3\} \times \{s_1, s_2\} \rightarrow \{p_2\}) = c_{\text{naturelle}}(\{p_3\} \times \{s_1, s_2\} \rightarrow \{p_2\}) = \frac{|\{c_1\}|}{|\{c_1\}|} = 1$. Pourtant, on observe que $c_{\text{exclusive}}(\{p_1\} \times \{s_3\} \rightarrow \{p_2\}) = c_{\text{naturelle}}(\{p_1\} \times \{s_3\} \rightarrow \{p_2\}) = \frac{|\emptyset|}{|\{c_1\}|} = 0$.

Les contre-exemples mentionnés ci-dessus impliquent des règles qui n'introduisent pas, dans leurs conclusions, d'autres dimensions que celles en prémisses. Or, en Section 3.3.1, il est expliqué que la sémantique de la confiance pour ce type de règle est intuitive (et donc respectée par les deux confiances que nous avons proposées). Il s'ensuit l'incompatibilité des axiomes d'augmentation et de transitivité avec une définition intuitive de la confiance.

3.5 Règle d'association canonique

Nous définissons maintenant un principe d'équivalence entre règles d'association et le concept de canonicité.

Définition 12 (Équivalence syntaxique) *$\forall \mathcal{D}' \subseteq \mathcal{D}$, les règles d'association $X \to Y$ et $X \to Z$ sur $\mathcal{D}'$ sont syntaxiquement équivalentes si et seulement si $X \sqcup Y = X \sqcup Z$.*

À partir des Définitions 8, 9 et 11, on démontre directement le lemme suivant.

Lemme 1 *Deux règles d'association syntaxiquement équivalentes ont même fréquence, même confiance exclusive et même confiance naturelle.*

Chaque règle d'association canonique représente sa classe d'équivalence syntaxique.

Définition 13 (Règle d'association canonique) *$\forall \mathcal{D}' \subseteq \mathcal{D}$, une règle d'association $X \to Y$ sur $\mathcal{D}'$ est canonique si et seulement si $\forall D^i \in \mathcal{D}, \pi_{D^i}(X) \cap \pi_{D^i}(Y) = \emptyset$.*

Toute collection complète de règles d'association satisfaisant des contraintes sur leurs fréquences et/ou confiances peut être résumée, sans perte d'information, à celles qui, parmi elles, sont canoniques. En effet, étant donné une règle d'association canonique $X \to Y$ dans la collection, le Lemme 1 permet d'affirmer la présence, dans la collection, de toutes les règles qui lui sont syntaxiquement équivalentes. De plus, les construire est facile : ce sont les règles d'association de la forme $X \to Y \sqcup Z$ avec $Z \sqsubseteq X$.

4 Calcul de règles a priori intéressantes

Face à une relation n-aire $\mathcal{R} \subseteq \times_{D^i \in \mathcal{D}} D^i$, nous voulons calculer des collections de règles a priori intéressantes, ce qui se traduit ici par le calcul de *toutes* les règles d'association canoniques :

- définies sur un sous-ensemble $\mathcal{D}' \subsetneq \mathcal{D}$;
- ayant une fréquence supérieure à un seuil $\mu \in [0; 1]$;
- ayant une confiance exclusive supérieure à un seuil $\beta_{\text{exclusive}} \in [0; 1]$;
- ayant une confiance naturelle supérieure à un seuil $\beta_{\text{naturelle}} \in [0; 1]$.

Plus formellement, une fois qu'un analyste a spécifié $\mathcal{D}' \subsetneq \mathcal{D}$ et les différents seuils (μ, $\beta_{\text{exclusive}}$ et $\beta_{\text{naturelle}}$), l'algorithme PINARD[1] doit calculer :

$$\{X \to Y \text{ canonique sur } \mathcal{D}' \mid \begin{cases} f(X \to Y) \geq \mu \\ c_{\text{exclusive}}(X \to Y) \geq \beta_{\text{exclusive}} \\ c_{\text{naturelle}}(X \to Y) \geq \beta_{\text{naturelle}} \end{cases} \}$$

Cette tâche sera effectuée en trois étapes : la construction du domaine support, l'extraction de l'ensemble des associations qui satisfont la contrainte de fréquence minimale, puis l'extraction des règles dont les confiances exclusive et naturelle dépassent les seuils choisis par l'analyste.

[1] PINARD Is N-ary Association Rule Discovery

4.1 Construction du domaine support

Le domaine support des règles d'association sur $\mathcal{D}'$ est $D^{\text{support}} = \times_{D^i \in \mathcal{D} \setminus \mathcal{D}'} D^i$. Soit $\mathcal{D}_A = \mathcal{D}' \cup D^{\text{support}}$. La relation $\mathcal{R}_A$ sur $\mathcal{D}_A$ est construite de la façon suivante :

$$\mathcal{R}_A = \{(e_1, e_2, \ldots, e_{|\mathcal{D}'|}, (e_{|\mathcal{D}'|+1}, \ldots, e_n)) \mid (e_1, e_2, \ldots, e_{|\mathcal{D}'|}, e_{|\mathcal{D}'|+1}, \ldots, e_n) \in \mathcal{R}\} \ .$$

4.2 Extraction des associations fréquentes

La fréquence d'une règle d'association sur $\mathcal{D}'$ est supérieure ou égale à μ si et seulement si l'union de sa prémisse et de sa conclusion est une association dont le support contient au moins $\alpha = \lceil \mu \times |D^{\text{support}}| \rceil$ éléments. L'extraction complète de telles associations ressemble au problème de l'extraction des itemsets fréquents dans une relation binaire. Cependant, il est doit être généralisé au contexte des relations n-aires. Un algorithme comme Data-Peeler [7] résout un problème assez proche : il impose la fermeture des associations alors que nous souhaitons ici lister toutes les associations fréquentes, qu'elles soient fermées ou non. Nous avons donc modifié Data-Peeler et ne présentons ici qu'une vision très abstraite de cette phase (voir [7] pour des détails).

Extraire toutes les associations A sur $\mathcal{D}'$ avec au moins α éléments dans son support peut s'exprimer comme le calcul de chaque association $A \times A^{\text{support}}$ sur $\mathcal{D}_A$ satisfaisant les quatre contraintes suivantes :

- $\mathcal{C}_{\text{sur-}\mathcal{D}'}(A \sqcup A^{\text{support}}) \equiv \forall D^i \in \mathcal{D}', \pi_{D^i}(A) \neq \emptyset$;
- $\mathcal{C}_{\text{connecté}}(A \sqcup A^{\text{support}}) \equiv A \sqcup A^{\text{support}} \subseteq \mathcal{R}_A$;
- $\mathcal{C}_{\text{support-entier}}(A \sqcup A^{\text{support}}) \equiv A^{\text{support}} = s(A)$;
- $\mathcal{C}_{\alpha\text{-fréquent}}(A \sqcup A^{\text{support}}) \equiv |A^{\text{support}}| \geq \alpha$.

La dernière contrainte traduit l'obligation, pour les règles utilisant tous les éléments de $\cup_{D^i \in \mathcal{D}'} \pi_{D^i}(A)$, d'excéder la fréquence minimale μ. En effet $\frac{|s(A)|}{|D^{\text{support}}|} \geq \mu$ équivaut à $|s(A)| \geq \alpha$ et, comme l'avant-dernière contrainte ($A^{\text{support}} = s(A)$) doit également être vérifiée, on trouve bien $|A^{\text{support}}| \geq \alpha$. L'avant-dernière contrainte, $\mathcal{C}_{\text{support-entier}}$, force un support « fermé ». En effet, par définition du support (Définition 2), ajouter un élément à A^{support} ($= s(A)$) viole forcément $\mathcal{C}_{\text{connecté}}$. $\mathcal{C}_{\text{support-entier}}(A \sqcup A^{\text{support}})$ équivaut ainsi à $\forall t \in D^{\text{support}} \setminus A^{\text{support}}$, $A \sqcup \{t\} \not\sqsubseteq \mathcal{R}_A$. C'est sous cette forme que nous l'utiliserons.

L'extracteur, que nous appelons Associations, parcourt l'espace de recherche en le partitionnant en deux à chaque appel récursif. L'énumération suit donc un arbre binaire. À chaque nœud de l'arbre, deux associations, appelées U et V, sont telles que U est la plus petite association (au sens de $\sqsubseteq$) qui pourra être extraite depuis ce nœud, $U \sqcup V$ la plus grande (au sens de $\sqsubseteq$). Ainsi, l'appel initial de Associations se fait avec $U = \emptyset$ et $V = \times_{D^i \in \mathcal{D}_A} D^i$ et toutes les associations dans $\mathcal{R}_A$ et vérifiant les quatre contraintes listées précédemment sont extraites. Les nœuds qui ne sont pas des feuilles ont deux fils. Un premier fils où un élément $e \in \cup_{D^i \in \mathcal{D}_A} \pi_{D^i}(V)$ est choisi pour être présent dans les associations qui seront extraites dans le sous-arbre d'énumération dont il est racine (e est « déplacé » de V vers U). Un second fils où ce même élément est déclaré absent des associations dans le sous-arbre d'énumération dont il est racine (e est « supprimé » de V).

Deux raisons peuvent faire qu'un nœud est une feuille de l'arbre d'énumération. La première raison est l'assurance qu'au moins une des quatre contraintes n'est vérifiée par

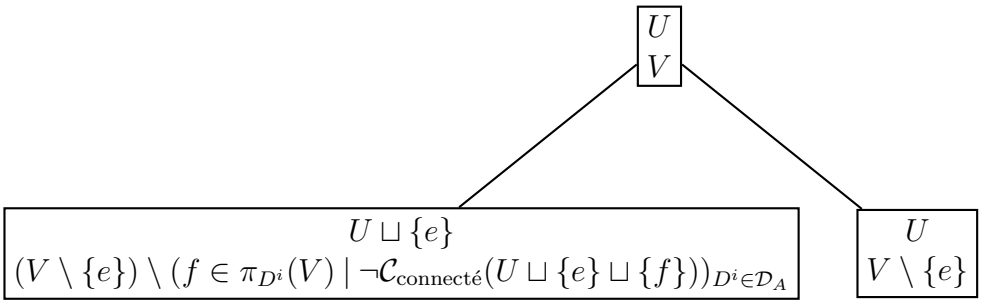


FIG. 1: Enumération de l'élément $e \in \cup_{D^i \in \mathcal{D}_A} \pi_{D^i}(V)$.

aucune association U dans le sous-arbre d'énumération qui dériverait du nœud. C'est le cas lorsque :

- $\exists D^i \in \mathcal{D}' \mid \pi_{D^i}(U \sqcup V) = \emptyset$ ($\mathcal{C}_{\text{sur-}\mathcal{D}'}$ est violée) ;
- $\forall D^i \in \mathcal{D}_A, \pi_{D^i}(U) \neq \emptyset \wedge U \not\subseteq \mathcal{R}_A$ ($\mathcal{C}_{\text{connecté}}$ est violée) ;
- $\exists t \in D^{\text{support}} \setminus \pi_{D^{\text{support}}}(U \sqcup V) \mid \Big((U \sqcup V) \setminus \pi_{D^{\text{support}}}(U \sqcup V)\Big) \sqcup \{t\} \subseteq \mathcal{R}_A$ ($\mathcal{C}_{\text{support-entier}}$ est violée) ;
- $|\pi_{D^{\text{support}}}(U \sqcup V)| < \alpha$ ($\mathcal{C}_{\alpha\text{-fréquent}}$ est violée).

Les preuves de ces propriétés d'élagage reposent sur une généralisation des notions de monotonie et d'anti-monotonie qui sont vérifiées par les quatre contraintes. La contrainte $\mathcal{C}_{\text{connecté}}$, dont la variable a été remplacée par U, est monotone : lorsque U viole la contrainte, toutes les associations plus grandes (au sens de $\sqsubseteq$) la violent également. De façon duale, les autres contraintes, dont la variable a été remplacée par $U \sqcup V$, sont anti-monotones : lorsque $U \sqcup V$ viole la contrainte, toutes les associations plus petites (au sens de $\sqsubseteq$) la violent également. L'autre raison qui peut faire qu'un nœud est une feuille de l'arbre d'énumération est que $V = \emptyset$. Il n'y a alors plus d'élément à énumérer. Si les quatre contraintes sont vérifiées, U est alors une association à partir de laquelle des règles d'association seront construites.

Une stratégie d'énumération améliorée évite de générer des nœuds violant $\mathcal{C}_{\text{connecté}}$ puis d'élaguer l'espace de recherche. À la place, à chaque appel récursif, on supprime de $\cup_{D^i \in \mathcal{D}_A} \pi_{D^i}(V)$ les éléments qui, si ils étaient « déplacés » vers U, violeraient $\mathcal{C}_{\text{connecté}}$. Ainsi, après avoir choisi un élément e à énumérer, les nœuds fils sont tels que décrits par Fig. 1. L'algorithme d'extraction des associations fréquentes est donné sous forme de pseudo-code (Algorithme 1).

4.3 Extraction des règles avec une confiance minimale

À partir d'une association fréquente A extraite par ASSOCIATIONS, il s'agit maintenant de construire des règles d'association canoniques utilisant *tous* les éléments de $\cup_{D^i \in \mathcal{D}'} \pi_{D^i}(A)$. Chacune de ces règles, $P \rightarrow C$, répartit ces éléments entre prémisse, P, et conclusion, C. En d'autres termes $P \sqcup C = A$. Pour énumérer ces règles, la stratégie d'énumération choisie construit un arbre. À chaque nœud de l'arbre est associée une règle d'association candidate. En d'autres termes, P et C sont instanciés et, si $P \rightarrow C$ vérifie les contraintes de confiances naturelle et exclusive minimales, alors elle est retenue.

En ne regardant que les conclusions, C, des règles (étant donné $A = P \sqcup C$, la

Entrée : (U, V)
Sortie : Toutes les associations fréquentes qui sont plus grandes que U et plus petites que $U \sqcup V$ (au sens de $\sqsubseteq$)
si $\mathcal{C}_{\text{sur-}\mathcal{D}'}(U \sqcup V) \wedge \mathcal{C}_{\text{support-entier}}(U \sqcup V) \wedge \mathcal{C}_{\alpha\text{-fréquent}}(U \sqcup V)$ **alors**
  si $V = \emptyset$ **alors**
    Sortir $U \setminus \pi_{D^{\text{support}}}(U)$
  sinon
    Choisir $e \in \cup_{D^i \in \mathcal{D}_A} \pi_{D^i}(V)$
    Associations$(U \sqcup \{e\},$
    $(V \setminus \{e\}) \setminus (f \in \pi_{D^i}(V) \mid \neg\mathcal{C}_{\text{connecté}}(U \sqcup \{e\} \sqcup \{f\}))_{D^i \in \mathcal{D}_A})$
    Associations$(U,\ V \setminus \{e\})$
  fin si
fin si

Algorithme 1 : Associations.

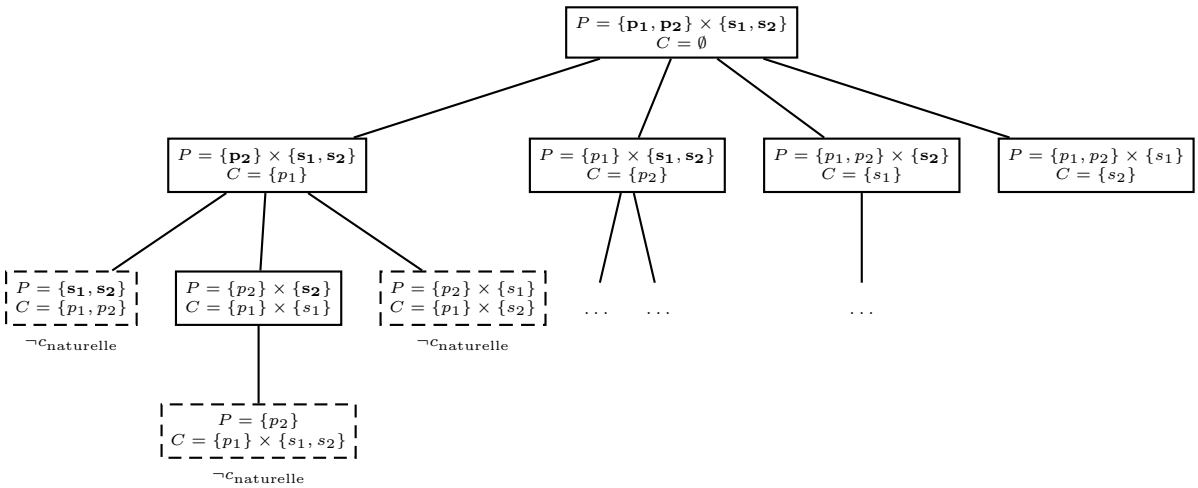


Fig. 2: Calcul des règles à partir d'une association.

prémisse, P, est unique), l'arbre qui est construit est, en fait, celui d'APriori [3]. En particulier, sa racine est $A \to \emptyset$ et C grandit d'un élément (via $\sqcup$) à chaque niveau de l'arbre (en parallèle, P se voit retirer ce même élément). Néanmoins cet arbre est, dans notre cas, parcouru en profondeur et ce n'est pas une fréquence minimale qui l'élague mais la confiance naturelle minimale. Le théorème à l'œuvre est donc le Théorème 2.

Par exemple, dans $\mathcal{R}_E$, considérons l'extraction de règles d'association canoniques ayant une confiance naturelle d'au moins $0,6$. Fig. 2 illustre le processus de production de ces règles à partir de l'association $A = \{p_1, p_2\} \times \{s_1, s_2\}$, extraite par Associations. Les éléments de $\cup_{D^i \in \mathcal{D}'} \pi_{D^i}(A)$ sont ordonnés de façon arbitraire. Dans cet exemple, l'ordre $\prec$ choisi est $p_1 \prec p_2 \prec s_1 \prec s_2$. À chaque nœud, les éléments qui peuvent augmenter (via $\sqcup$) la conclusion sont ceux qui sont plus grands (selon $\prec$) que tous les éléments déjà en conclusion (autrement dit, plus grand que $\max_{\prec}(C)$, (pour l'appel initial à Associations, l'élément $\max_{\prec}(\emptyset)$ est défini comme plus petit que tous les autres dans l'ordre $\prec$). Sur la figure, ces éléments sont en gras. Un nœud sans élément en gras n'a aucun fils. Un nœud qui ne satisfait pas la contrainte de confiance naturelle minimale

(il est, sur la figure, encadré de pointillés), n'en a pas non plus. D'après le Théorème 2, cet élagage est *sûr* : aucune règle avec une confiance suffisante n'est manquée.

Comme nous l'avons vu, et contrairement à la confiance naturelle, la confiance exclusive n'est pas anti-monotone. Vérifier la contrainte de confiance exclusive minimale est donc l'ultime condition à vérifier pour produire la règle mais elle ne donne jamais lieu à élagage. L'Algorithme 2 résume l'extraction des règles canoniques de confiance suffisante depuis une association fréquente A. Pour obtenir de bonnes performances, précisons que les confiances (exclusive et naturelle) sont, autant que possible, calculées sans retour à $\mathcal{R}_A$. Déjà $|s(P \sqcup C)| = |s(A)|$, qui intervient dans les deux définitions, est constante et connue dès l'extraction de A par ASSOCIATIONS. Ensuite $|s(P)|$ est connue si aucune de ses composantes n'est vide : en effet, puisque $P \sqsubseteq (P \sqcup C)$, le Théorème 1 nous assure que P est une association fréquente sur $\mathcal{D}'$ et a donc été extraite par ASSOCIATIONS. Enfin, à chaque calcul de $|s(P)|$ (P a alors une composante vide) ou de $|s_{\mathcal{D}'}(P)|$ depuis $\mathcal{R}_A$, la valeur est stockée pour éviter de la calculer à nouveau si cette même prémisse est considérée pour une autre règle.

Entrée : (P, C)
Sortie : Toutes les règles d'association canoniques utilisant tous les éléments de $\cup_{D^i \in \mathcal{D}'} \pi_{D^i}(P \sqcup C)$, avec une prémisse plus petite que P (selon $\sqsubseteq$), une conclusion plus grande que C (selon $\sqsubseteq$) et satisfaisant les contraintes de confiance minimale
pour tout $e \succ \max_{\prec}(C)$ **faire**
  $(P', C') \leftarrow (P \setminus \{e\}, C \sqcup \{e\})$
  si $c_{\text{naturelle}}(P' \rightarrow C') \geq \beta_{\text{naturelle}}$ **alors**
    si $c_{\text{exclusive}}(P' \rightarrow C') \geq \beta_{\text{exclusive}}$ **alors**
      Sortir $P' \rightarrow C'$
    fin si
    RÈGLES(P', C')
  fin si
fin pour

Algorithme 2 : RÈGLES.

Nous pouvons maintenant donner l'Algorithme 3 qui répond au problème du calcul des règles a priori intéressantes formalisé au début de la Section 4.

5 Validation empirique

Les expériences suivantes ont été effectuées sur un ordinateur Intel® Pentium® 4 avec un processeur cadencé à 3 GHz, 1 Go de RAM et le système d'exploitation GNU/Linux™. PINARD est codé en C++ et compilé avec GCC 4.2.4. Nous donnons quelques éléments qualitatifs montrant que l'algorithme proposé permet de calculer des motifs effectivement pertinents dans des données réelles. Nous considérons ensuite quelques éléments davantage quantitatifs.

Entrée : Relation $\mathcal{R}$ sur $\mathcal{D} = \{D^1, \ldots, D^n\}$, $\mathcal{D}' \subsetneq \mathcal{D}$, $(\mu, \beta_{\text{exclusive}}, \beta_{\text{naturelle}}) \in [0;1]^3$
Sortie : Toutes les règles d'association canoniques sur $\mathcal{D}'$ satisfaisant les contraintes de fréquence et de confiances minimales
$D^{\text{support}} \leftarrow \times_{D^i \in \mathcal{D} \setminus \mathcal{D}'} D^i$
$(\mathcal{D}_A, \mathcal{R}_A) \leftarrow (\mathcal{D}' \cup D^{\text{support}}, \emptyset)$
pour tout $(e_1, e_2, \ldots, e_{|\mathcal{D}'|}, e_{|\mathcal{D}'|+1}, \ldots, e_n) \in \mathcal{R}$ **faire**
  $\mathcal{R}_A \leftarrow \mathcal{R}_A \cup (e_1, e_2, \ldots, e_{|\mathcal{D}'|}, (e_{|\mathcal{D}'|+1}, \ldots, e_n))$
fin pour
$\alpha \leftarrow \lceil \mu \times |D^{\text{support}}| \rceil$
$\mathcal{A} \leftarrow$ ASSOCIATIONS$(\emptyset, \times_{D^i \in \mathcal{D}_A} D^i)$
pour tout $A \in \mathcal{A}$ **faire**
  RÈGLES$(A, \emptyset)$
fin pour

Algorithme 3 : PINARD.

5.1 Étude qualitative

`DistroWatch` est un site Web qui rassemble une information complète sur les distributions GNU/Linux, BSD et Solaris. Chaque distribution est décrite sur une page séparée. Lorsque qu'un visiteur charge une page, on considère que la distribution qu'elle décrit l'intéresse. L'adresse IP du visiteur nous permet de connaître son pays. Les données produites de l'année 2004 à l'année 2007 sont agrégées par semestre, par pays et par distribution. Seuls les pays associés à au moins 2000 consultations d'une distribution lors d'un semestre, ont été gardés. Les données numériques sont ensuite normalisées de sorte que tous les pays (resp. tous les semestres) aient la même importance. Enfin, elles sont transformées en une relation ternaire listant les triplets les plus significatifs. Ces derniers sont choisis à l'aide d'une procédure locale (i.e., par distribution) inspirée du calcul d'une valeur p : pour chaque distribution, on garde ses triplets associés aux plus grandes valeurs numériques jusqu'à ce que leur somme atteigne 25% de la somme de toutes les valeurs impliquant la distribution. Nous appelons la relation ainsi obtenue $\mathcal{R}_{\texttt{DistroWatch}}$. Elle contient 16499 triplets impliquant 8 semestres, 87 pays et 555 distributions.

Nous souhaitons découvrir des règles associant pays et distributions (ces deux dimensions forment l'ensemble que nous avons appelé $\mathcal{D}'$ jusqu'à maintenant). PINARD est utilisé avec pour seuils de fréquence et de confiances $\mu = 0,8$, $\beta_{\text{exclusive}} = 0,8$ et $\beta_{\text{naturelle}} = 0,8$. On extrait alors 66 règles d'association canoniques. Parmi elles :

- ◇ {Bayanihan} → {Philippines} (f : 1, $c_{\text{naturelle}}$: 1, $c_{\text{exclusive}}$: 0,8) ;
- ◇ {blackPanther, Frugalware} → {Hongrie} (f :0,875, $c_{\text{naturelle}}$: 1, $c_{\text{exclusive}}$:1) ;
- ◇ {Momonga, Plamo} → {Japon} × {Berry} (f : 0,875, $c_{\text{naturelle}}$: 1, $c_{\text{exclusive}}$: 1) ;
- ◇ {Japon} × {Plamo} → {Berry, Omoikane} (f : 0,875, $c_{\text{naturelle}}$: 0,875, $c_{\text{exclusive}}$: 0,875).

La distribution *Bayanihan* est développée aux Philippines. C'est pourquoi elle est plus consultée par les internautes de ce pays que par ceux des autres pays. En effet, la règle {Bayanihan} → {Philippines} (f : 1, $c_{\text{naturelle}}$: 1, $c_{\text{exclusive}}$: 0,8) indique que *Bayanihan* intéresse particulièrement l'internaute Philippin, quel que soit le semestre (la

fréquence de la règle vaut 1). De plus, *Bayanihan* intéresse peu en dehors des Philippines (sa confiance exclusive est proche de 1). La même analyse peut-être faite pour la règle {blackPanther, Frugalware} → {Hongrie} (f :0,875, $c_{\text{naturelle}}$: 1, $c_{\text{exclusive}}$:1). En effet, ce sont des Hongrois qui concoctent les distributions *blackPanther* et *Frugalware*. Comme on a une conjonction de distributions hongroises en prémisse, la confiance exclusive est encore plus forte que celle de la règle précédente. Elle vaut 1 ce qui signifie que jamais ces deux distributions sont, le même semestre, fréquemment consultées ensemble en dehors de Hongrie. Le pays d'origine des distributions *Momonga*, *Plamo*, *Berry* et *Omoikane* est le Japon. La règle {Momonga,Plamo} → {Japon} × {Berry} (f : 0,875, $c_{\text{naturelle}}$: 1, $c_{\text{exclusive}}$: 1) signifie qu'à l'extérieur des frontières nippones, les pages décrivant *Momonga* et *Plamo* ne sont jamais toutes deux fréquemment chargées un même semestre ($c_{\text{exclusive}} = 1$). De plus, lorsque *Momonga*, *Plamo* sont consultées, l'internaute japonais visite aussi *Berry* ($c_{\text{naturelle}} = 1$). La règle {Japon} × {Plamo} → {Berry, Omoikane} (f : 0,875, $c_{\text{naturelle}}$: 0,875, $c_{\text{exclusive}}$: 0,875) indique que, la majorité du temps, le Japonais qui visite *Plamo* visite aussi *Berry* et *Omoikane*.

Les règles que nous venons de détailler font sens puisque les distributions développées spécifiquement par et pour un pays, intéressent particulièrement les internautes de ce pays. Il se trouve que les règles que nous n'avons pas discutées mais qui ont des distributions en prémisse et des pays en conclusion, sont majoritairement interprétables de cette façon. Pour en rendre compte et valider les mesures de confiances que nous avons définies, la Fig. 3 représente, pour différents paramétrages, la valeur suivante :

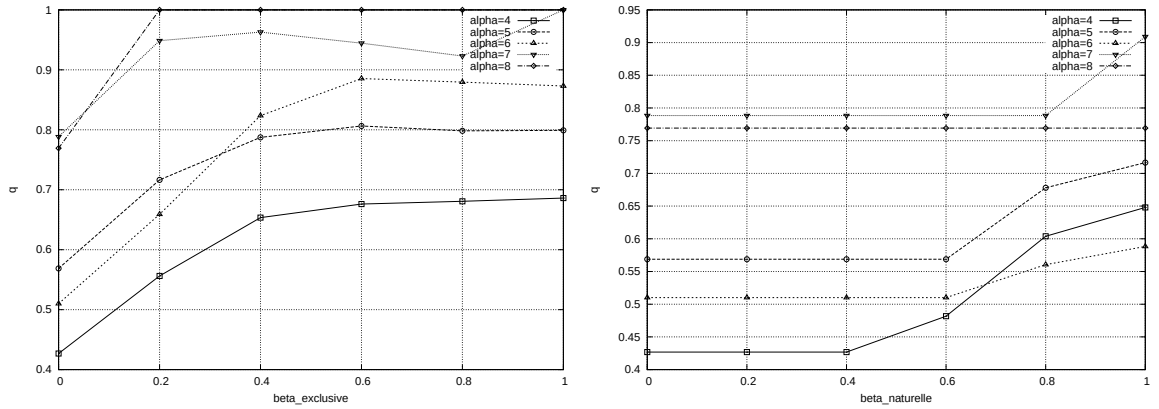
$$q = \frac{|\{D \to \{p\} \mid D \subseteq D^{\text{distributions}} \wedge p \in D^{\text{pays}} \wedge (\exists d \in D \text{ tel que origine}(d) = p)\}|}{|\{D \to P \mid D \subseteq D^{\text{distributions}} \wedge P \subseteq D^{\text{pays}}\}|}$$

où origine(d) est le pays d'où provient la distribution d. Lorsque les seuils de confiances minimales augmentent, q augmente. Globalement, lorsque le seuil de fréquence minimale augmente, q augmente aussi. Cela corrobore donc empiriquement les choix des sémantiques associées à ces mesures. Avec une confiance naturelle, les irrégularités dans l'augmentation de q avec le seuil de fréquence s'explique par des collections assez réduites de règles ayant des distributions en prémisse et des pays en conclusion. En revanche, ce type de règles atteint facilement une forte confiance exclusive (alors que les règles suivant un autre « schéma » n'ont pas de raison d'être exclusives). D'ailleurs q augmente plus vite avec $c_{\text{exclusive}}$ qu'avec $c_{\text{naturelle}}$. Les paliers observés sur la Fig. 3b pour $\beta_{\text{naturelle}} \leq \mu$ sont des conséquences directes des Définitions 8 et 11 : les règles extraites sont les mêmes.

5.2 Élagage dans la génération des règles

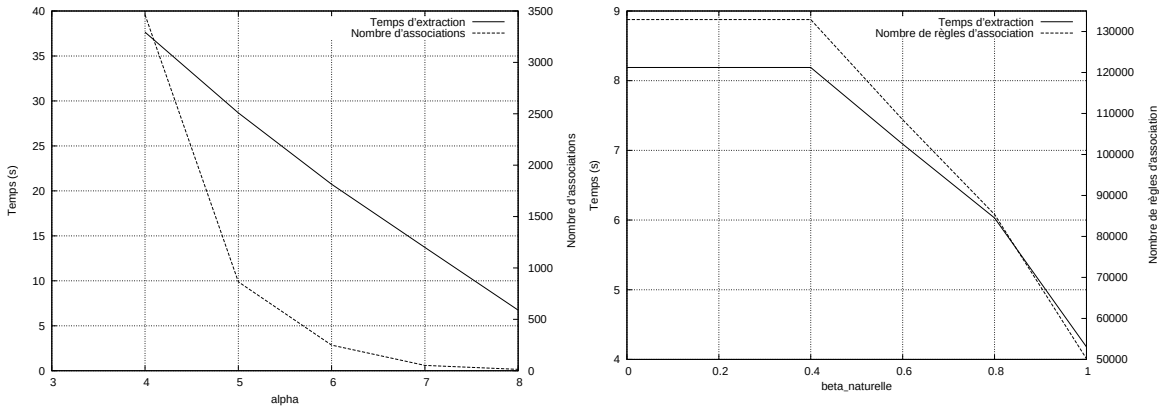
Si le seuil de fréquence augmente, le nombre d'associations fréquentes ne peut que diminuer. Dans l'algorithme Associations, la vérification de la contrainte $\mathcal{C}_{\alpha\text{-fréquent}}$ va donc élaguer de plus grandes régions de l'espace de recherche où la contrainte est violée. Le temps d'extraction des associations va donc décroître avec le seuil de fréquence minimale. Fig. 4a illustre l'efficacité de l'élagage de l'algorithme Associations sur $\mathcal{R}_{\texttt{DistroWatch}}$.

En exploitant le Théorème 2, l'algorithme Règles élague les arbres dérivant des associations fréquentes. Ainsi, quand le seuil de confiance naturelle augmente, le nombre



(a) q lorsque $\beta_{\text{exclusive}}$ varie ($\beta_{\text{naturelle}} = 0$). (b) q lorsque $\beta_{\text{naturelle}}$ varie ($\beta_{\text{exclusive}} = 0$).

FIG. 3: Validation qualitative de la confiance



(a) Extraction des associations. (b) Extraction des règles d'association.

FIG. 4: Efficacité de l'élagage

de règles et le temps de calcul de ces règles diminuent (Fig. 4b). Cette expérience est réalisée sur $\mathcal{R}_{\texttt{DistroWatch}}$ avec $\beta_{exclusive} = 0$ et à partir des règle ayant une fréquence supérieure ou égale à $0,5$. Le seuil de confiance naturelle varie de 0 à 1.

6 Travaux connexes

Depuis la proposition initiale du problème dans [2] et son raffinement dans [3], le calcul de règles d'associations ayant une fréquence et une confiance suffisante dans des relations binaires éventuellement très grandes a été intensivement étudié. Les généralisations de cette tâche prototypique ont été nombreuses. Certains auteurs ont cherché à caractériser les types de règles selon le nombre de dimensions présentes dans une règle d'association. Ainsi, [22] fait la différence entre les règles d'association intra dimensionnelles, les règles d'association inter dimensionnelles, et les règles hybrides. Dans le cas des règles d'association intra dimensionnelles, tous les éléments de la règle n'apparti-

ennent qu'à une seule dimension et le cas des relations avec deux dimensions (disons $Transactions \times Items$ avec des règles sur les $Items$) est donc particulièrement maîtrisé. En 2006, Schmitz et al. ont proposé dans [19] le calcul de règles d'association intra dimensionnelles dans une relation n-aire ($n \geq 2$). Ils recherchent des règles d'association sur chaque dimension de la relation et ils considèrent tous les uplets qui appartiennent au produit cartésien de toutes les autres dimensions comme un domaine de transactions.

Les règles d'association inter dimensionnelles ont été proposées pour permettre de trouver des associations ou co-occurrences entre plusieurs dimensions de la relation. Il s'agit de découvrir des règles d'association dont les éléments appartiennent à quelques dimensions distinctes et où aucune dimension n'est répétée (i.e., dans la règle, il n'existe pas deux éléments qui appartiennent à une même dimension) [14, 17]. Dans ces propositions, l'extraction des règles inter dimensionnelles est guidée par des méta-règles ou gabarits. Une méta-règle est un schéma de règle avec quelques prédicats distincts (aucun prédicat n'est répété). L'absence de répétitions est une limitation à l'expressivité des règles. D'autres auteurs ont proposé des algorithmes ad-hoc d'extraction de règles hybrides permettant la répétition de quelques dimensions [22, 10, 21]. Dans notre étude, nous fouillons, à partir de la relation n-aire ($n \geq 2$), des règles d'association avec aucune des restrictions évoquées ci-dessus. Les éléments d'une règle appartiennent à n'importe quelles dimensions et, pour chacune de ces dimensions, on peut avoir un ou plusieurs éléments. De plus, la répartition entre prémisse et conclusion ne souffre, là encore, d'aucune contrainte.

7 Conclusion

Nous considérons qu'il est important de développer de nouvelles tâches de fouille de données dans des relations d'arité arbitraire. Dans cet article, nous avons étudié la généralisation du concept de règle d'association et montré que l'étude des classiques mesures d'intérêt objectives que sont la fréquence et la confiance était délicate dans un contexte de relation n-aire avec $n > 2$. Nous avons donc formalisé une généralisation pour laquelle les prémisses et les conclusions des règles peuvent impliquer des sous-ensembles de n'importe lesquelles des n dimensions. Nous avons fixé la sémantique de telles règles via des mesures de confiance complémentaires appelées confiance naturelle et confiance exclusive. Un premier algorithme d'extraction a été proposé et implémenté. Notre validation empirique a permis d'illustrer la découverte de motifs effectivement pertinents dans des données réelles.

Les perspectives de ce travail sont multiples. L'une des plus importantes consiste à concevoir des contraintes primitives pour qualifier des règles d'association pertinentes et bien sûr à savoir les exploiter au mieux pour avoir des extractions faisables en pratique. Ceci peut nous conduire vers l'étude de nouvelles mesures de la qualité puisque l'on connaît déjà bien les limites des mesures de fréquence et de confiance et que bien d'autres propositions existent pour le cas binaire (voir, e.g., [15]). Une autre perspective est d'étudier la sémantique des règles dans le cas de relations n-aires particulières. Par exemple, nous étudions actuellement les applications possibles du codage de graphes dynamiques dans des relations ternaires et nous voulons comprendre l'impact des règles multi-dimensionnelles dans ce contexte important pour ses perspectives d'applications.

Remerciements. Ce travail est partiellement financé sur le contrat ANR MDCO Bingo2 (2007-2010) et par une bourse du gouvernement du Vietnam.

Références

[1] M. Agier, J.-M. Petit, and E. Suzuki. Unifying framework for rule semantics : Application to gene expression data. *Fundamenta Informaticae*, 78(4) :543–559, 2007.

[2] R. Agrawal, T. Imielinski, and A. N. Swami. Mining association rules between sets of items in large databases. In *SIGMOD '93 : Int. Conf. on Management of Data*, pages 207–216. ACM Press, 1993.

[3] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A. I. Verkamo. Fast discovery of association rules. In *Advances in Knowledge Discovery and Data Mining*, pages 307–328. AAAI/MIT Press, 1996.

[4] W. W. Armstrong. Dependency structures of data base relationships. In *IFIP '74 : Information Processing Congress*, pages 580–583, 1974.

[5] J.-F. Boulicaut, L. De Raedt, and H. Mannila, editors. *Constraint-based Mining and Inductive Databases*, volume 3848 of *LNCS*. Springer, 2006.

[6] L. Cerf, J. Besson, C. Robardet, and J.-F. Boulicaut. Data-Peeler : Constraint-based closed pattern mining in n-ary relations. In *SDM '08 : 8th SIAM Int. Conf. on Data Mining*, pages 37–48. SIAM, 2008.

[7] L. Cerf, J. Besson, C. Robardet, and J.-F. Boulicaut. Closed patterns meet n-ary relations. *ACM Trans. on Knowledge Discovery from Data*, 3(1) :1–36, 2009.

[8] L. Dehaspe and L. De Raedt. Mining association rules in multiple relations. In *ILP '97 : 7th Int. Workshop on Inductive Logic Programming*, pages 125–132. Springer, 1997.

[9] B. Ganter, G. Stumme, and R. Wille. *Formal Concept Analysis, Foundations and Applications*, volume 3626 of *LNCS*. Springer, 2005.

[10] T. Imielinski, L. Khachiyan, and A. Abdulghani. Cubegrades : Generalizing association rules. *Data Mining and Knowledge Discovery*, 6(3) :219–257, 2002.

[11] R. Jaschke, A. Hotho, C. Schmitz, B. Ganter, and G. Stumme. Trias–an algorithm for mining iceberg tri-lattices. In *ICDM '06 : 6th Int. Conf. on Data Mining*, pages 907–911. IEEE Computer Society, 2006.

[12] T.-Y. Jen, D. Laurent, N. Spyratos, and O. Sy. Towards mining frequent queries in star schemes. In *Knowledge Discovery in Inductive Databases, 4th Int. Workshop KDID '05, Revised Selected and Invited Papers*, pages 104–123. Springer, 2005.

[13] L. Ji, K.-L. Tan, and A. K. H. Tung. Mining frequent closed cubes in 3D data sets. In *VLDB '06 : 32nd Int. Conf. on Very Large Data Bases*, pages 811–822. VLDB Endowment, 2006.

[14] M. Kamber, J. Han, and J. Y. Chiang. Metarule-guided mining of multi-dimensional association rules using data cubes. In *KDD '97 : 3rd SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, pages 207–210. AAAI Press, 1997.

[15] P. Lenca, P. Meyer, B. Vaillant, and S. Lallich. On selecting interestingness measures for association rules : User oriented description and multiple criteria decision aid. *European Journal of Operational Research*, 184(2) :610–626, 2008.

[16] H. Mannila and H. Toivonen. Multiple uses of frequent sets and condensed representations. In *KDD '96 : 2nd SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, pages 189–194. AAAI Press, 1996.

[17] R. Ben Messaoud, S. Loudcher Rabaséda, O. Boussaid, and R. Missaoui. Enhanced mining of association rules from data cubes. In *DOLAP '06 : 9th ACM Int. Workshop on Data Warehousing and OLAP*, pages 11–18, 2006.

[18] F. Rioult, B. Zanuttini, and B. Crémilleux. Apport de la négation pour la classification supervisée à l'aide d'associations. In *CAp '08 : Actes de la 10ème Conférence d'Apprentissage*, pages 183–196. Cépaduès éditions, 2008.

[19] C. Schmitz, A. Hotho, R. Jäschke, and G. Stumme. Mining association rules in folksonomies. In *Data Science and Classification*, pages 261–270. Springer, 2006.

[20] G. Stumme, R. Taouil, Y. Bastide, N. Pasqier, and L. Lakhal. Computing iceberg concept lattices with TITANIC. *Data & Knowledge Engineering*, 42(65) :189–222, 2002.

[21] H. Cokrowijoyo Tjioe and D. Taniar. Mining association rules in data warehouses. *Int. Journal of Data Warehousing and Mining*, 1(3) :28–62, 2005.

[22] H. Zhu. On-line analytical mining of association rules. Master's thesis, Simon Fraser University, Burnaby, British Columbia, Canada, 1998.

Annexe

Preuve du Théorème 1 : D'après les Définitions 6 et 2, on a :

- $X \sqsubseteq Y \Rightarrow \begin{cases} \mathcal{D}_X \subseteq \mathcal{D}_Y \\ \forall D^i \in \mathcal{D}, \pi_{D^i}(X) \subseteq \pi_{D^i}(Y) \end{cases}$;
- $s(Y) = \{t \in \times_{D^i \in \mathcal{D} \setminus \mathcal{D}_Y} D^i \mid \forall y \in Y, y \cdot t \in \mathcal{R}\}$;
- $s(X) = \{w \in \times_{D^i \in \mathcal{D} \setminus \mathcal{D}_X} D^i \mid \forall x \in X, x \cdot w \in \mathcal{R}\}$
  $= \{u \cdot t \mid u \in \times_{D^i \in \mathcal{D}_Y \setminus \mathcal{D}_X} D^i, t \in \times_{D^i \in \mathcal{D} \setminus \mathcal{D}_Y} D^i \text{ et } \forall x \in X, x \cdot u \cdot t \in \mathcal{R}\}$.

Soit $\pi_{\mathcal{D} \setminus \mathcal{D}_Y} s(X) = \{t \in \times_{D^i \in \mathcal{D} \setminus \mathcal{D}_Y} D^i \mid \exists u \in \times_{D^i \in \mathcal{D}_Y \setminus \mathcal{D}_X} D^i \text{ tel que } \forall x \in X, x \cdot u \cdot t \in \mathcal{R}\}$.

On a $\begin{cases} s(Y) \subseteq \pi_{\mathcal{D} \setminus \mathcal{D}_Y} s(X) \\ |\pi_{\mathcal{D} \setminus \mathcal{D}_Y} s(X)| \leq |s(X)| \end{cases} \Rightarrow |s(Y)| \leq |\pi_{\mathcal{D} \setminus \mathcal{D}_Y} s(X)| \leq |s(X)|$.

Preuve du Théorème 2 : D'après la Définition 10, on a :

- $s_{\mathcal{D} \setminus \mathcal{D}'}(X) = \{t \in \times_{D^i \in \mathcal{D} \setminus \mathcal{D}'} D^i \mid \exists u \in \times_{D^i \in \mathcal{D}' \setminus \mathcal{D}_X} D^i \text{ tel que } \forall x \in X, x \cdot u \cdot t \in \mathcal{R}\}$;
- $s_{\mathcal{D} \setminus \mathcal{D}'}(X') = \{t \in \times_{D^i \in \mathcal{D} \setminus \mathcal{D}'} D^i \mid \exists u' \in \times_{D^i \in \mathcal{D}' \setminus \mathcal{D}_{X'}} D^i \text{ tel que } \forall x' \in X', x' \cdot u' \cdot t \in \mathcal{R}\}$.

Par ailleurs, comme $X \sqsubseteq X' \sqsubseteq Y$ et d'après la Définition 11 :

$$\begin{cases} s_{\mathcal{D} \setminus \mathcal{D}'}(X') \subseteq s_{\mathcal{D} \setminus \mathcal{D}'}(X) \\ c_{\text{naturelle}}(X \to Y \setminus X) = \frac{|s(Y)|}{|s_{\mathcal{D} \setminus \mathcal{D}'}(X)|} \\ c_{\text{naturelle}}(X' \to Y \setminus X') = \frac{|s(Y)|}{|s_{\mathcal{D} \setminus \mathcal{D}'}(X')|} \end{cases}$$

$\Rightarrow c_{\text{naturelle}}(X \to Y \setminus X) \leq c_{\text{naturelle}}(X' \to Y \setminus X')$.