Optimisation d'Extraction de Motifs : une nouvelle Méthode fondée sur la Transposition de Données

François RIOULT et Bruno CRÉMILLEUX

GREYC, CNRS - UMR 6072, Université de Caen
F-14032 Caen Cédex France
{Francois.Rioult,Bruno.Cremilleux}@info.unicaen.fr

Résumé : Chercher dans une base de données les motifs d'attributs (les colonnes) ou les groupes d'objets (les lignes) qui vérifient certaines propriétés est une tâche classique de la fouille de données, aujourd'hui bien maîtrisée. Néanmoins, certains contextes difficiles comme les données issues de la bio-informatique restent impraticables, en raison d'un nombre d'attributs disproportionné devant le nombre d'objets.

Dans ces conditions, il est naturellement tentant de transposer la matrice des données pour y pratiquer plus efficacement l'extraction des motifs. Cet article expose cette nouvelle méthode et montre son intérêt mais aussi les difficultés à résoudre pour que cette approche soit fructueuse. En tirant profit de la connexion de Galois, l'extraction réalisée dans la base transposée permet d'inférer des résultats sur les données initiales. Nous montrons les apports de cette pratique sur des données contenant un grand nombre d'attributs, comme les données de génome, ainsi que son application possible à l'extraction sous contrainte monotone et l'obtention de la totalité de l'ensemble des motifs fermés.

Mots-clés : extraction de motifs fermés, treillis de concepts, connexion de Galois, transposition de données.

1 Introduction

L'extraction des motifs d'une base de données est aujourd'hui une tâche bien maîtrisée. Depuis la création du premier algorithme APRIORI (Agrawal *et al.*, 1996), les techniques se sont affinées et on dispose maintenant d'autres outils comme les représentations condensées et les motifs fermés (Mannila & Toivonen, 1996; Bastide *et al.*, 2002). Le traitement des bases de données dont la configuration classique comporte un grand nombre de lignes (plusieurs millions) devant le nombre de colonnes (quelques milliers au maximum) n'est plus un souci. Les motifs découverts sont constitués des attributs correspondant aux colonnes.

En revanche, certaines données issues de contextes spécifiques comme la biologie et le séquençage des gènes n'ont pas encore trouvé d'algorithme permettant d'en extraire

efficacement de la connaissance. C'est le cas des bases comportant peu de lignes pour une grande quantité de colonnes (lors du séquençage de gènes, les lignes sont souvent les expériences réalisées et les colonnes les expressions de gène). Dans cette situation, il est tentant d'appliquer les techniques de fouille de données à la matrice transposée des données. La nouvelle base comporte alors des dimensions compatibles avec une extraction de motifs sur de nombreuses lignes (les gènes) et peu de colonnes (les expériences). Hélas, les résultats obtenus sont relatifs à des motifs d'expériences, ce qui passionne peu les biologistes. La difficulté du calcul est réduite, mais d'autres problèmes surgissent : interprétation du résultat, conversion des paramètres d'extraction, etc.

Nous nous proposons dans cet article d'étudier une nouvelle méthode d'extraction qui tire pleinement parti des caractéristiques géométriques de la base. Nous étudions l'extraction de l'information depuis la transposée de la base de données, puis utilisons la connexion de Galois pour inférer les résultats obtenus dans la base initiale. Cette connexion définit des concepts sous la forme d'une association unique entre un motif d'attributs et les objets qui les contiennent, ou en transposant, entre un groupe d'objets et les attributs qu'ils partagent. Quand le format est propice (peu d'objets, beaucoup d'attributs), la transposition de matrice permet de travailler sur des données dont l'exploration sera facilitée par un renversement des tendances lignes / colonne.

Nous pensons que cette méthode est nouvelle, même si elle intègre des outils classiques de la communauté apprentissage (extraction de motifs, connexion de Galois, treillis de concepts). Sa nouveauté réside dans les solutions qu'elle fournit à des problèmes jusqu'ici inaccessibles, et ceci par une combinaison astucieuse de méthodes conventionnelles dans un contexte de fouille de dimensions particulières.

Nous commençons par décrire les fondements théoriques de la recherche des motifs satisfaisant une propriété donnée dans la base, et en particulier la connexion de Galois. Puis nous définissons la transposition d'une extraction de motifs et montrons comment exploiter la connexion pour inférer des résultats sur la base originale. La dernière section est dédiée à l'étude des développements possibles de cette méthode : extraction de l'intégralité des motifs fermés, exploitation de contrainte monotone et de la bordure commune, et applications sur des données biologiques.

2 Extraction de motifs

Dans cette section sont rappelés les fondements théoriques qui nous permettent de développer la méthode d'extraction de motifs basée sur la transposition. Nous décrivons pour cela le treillis constitué par les motifs et la relation de spécialisation qui les relie, suivant le cadre de Mitchell (Mitchell, 1980). Puis nous détaillons l'apport des contraintes anti-monotones dans les algorithmes de recherche et précisons les bordures de théorie qui permettent de comparer les extractions, dans le cadre de Mannila et Toivonen (Mannila & Toivonen, 1997). Nous finissons avec la connexion de Galois, nécessaire à la compréhension de notre méthode, sur laquelle nous appuyons le reste de notre exposé.

	Attributs									
Objets	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}
o_1	1	1	1	1	0	1	1	0	0	0
o_2	1	1	1	1	0	0	0	0	1	1
o_3	1	1	1	1	0	0	0	0	1	1
o_4	0	0	0	0	1	1	1	1	1	1
o_5	1	0	1	0	1	1	1	1	0	0

TAB. 1 – Exemple de base de données

2.1 Treillis des motifs

Le tableau 1 présente une table qui constitue une base de données pour cinq objets d'étude (en lignes) et indique pour chacun l'absence ou la présence de 10 attributs binaires (en colonnes). Cet exemple sera repris tout du long de notre présentation. On remarque entre autres que le groupe d'objets o_1, o_2, o_3 partage les mêmes attributs a_1, a_2, a_3, a_4, et que o_4 et o_5 partagent a_5, a_6, a_7, a_8, etc. Nous utiliserons plus précisément le terme de motif d'attributs ou d'objets plutôt que de groupe.

Soit $\mathcal{O}$ une liste d'objets et $\mathcal{A}$ une liste d'attributs. Sur notre exemple, $\mathcal{O} = \{o_1, o_2, \dots o_5\}$ et $\mathcal{A} = \{a_1, a_2, \dots a_{10}\}$. Les données à explorer sont représentées par la matrice de la relation binaire $R \subset \mathcal{O} \times \mathcal{A}$ définie entre chaque objet et chaque attribut (cf. Table 1). Ainsi, $(o_i, a_j) \in R$ signifie que l'objet i porte l'attribut j. Une base de données bd est un triplet $(\mathcal{O}, \mathcal{A}, R)$ associant deux ensembles d'objets et d'attributs à l'aide d'une relation binaire.

L'ensemble $\mathcal{A}$ des attributs permet de construire le langage $\mathcal{L}_\mathcal{A} = 2^\mathcal{A}$ des motifs constitués d'éléments de $\mathcal{A}$. Nous cherchons alors parmi $\mathcal{L}_\mathcal{A}$ les motifs qui vérifient dans bd une propriété donnée q. Par exemple, connaître les motifs d'attributs rencontrés fréquemment (quand le nombre d'objets contenant ce motif, appelé *support*, dépasse un certain seuil). Sur notre base exemple, les motifs d'attributs $\{a_1, a_2, a_3, a_4\}$ et $\{a_9, a_{10}\}$ sont présents au moins trois fois, à la différence de $\{a_2, a_8\}$, qui n'est jamais présent.

L'ensemble $\mathcal{L}_\mathcal{A}$ des motifs d'attributs se représente naturellement sous forme d'un treillis (cf. Figure 1). En haut, les motifs de longueur 1 : $\{a_1\} \dots \{a_{10}\}$. Sur le niveau suivant, les motifs de longueur 2 : $\{a_1, a_2\}$, $\{a_1, a_3\}$, $\{a_1, a_4\}$, … $\{a_1, a_{10}\}$, $\{a_2, a_3\}$, … $\{a_9, a_{10}\}$. Puis les motifs de longueur 3, etc. L'avant dernier niveau contient les dix motifs de longueur 9 et le dernier contient l'unique motif de longueur 10.

Le treillis est orienté du haut (les motifs de longueur minimale, les singletons) vers le bas (les plus longs motifs possibles) suivant une relation de *spécialisation* (cadre de Mitchell (Mitchell, 1980)), qui fournit une méthode pour leur production à partir des singletons. Un motif X, constitué d'attributs de $\mathcal{A}$, est plus spécifique qu'un autre motif Y si $Y \subset X$. Le treillis peut alors être généré méthodiquement, niveau par niveau, des motifs les plus courts aux plus longs, à l'aide d'une spécialisation progressive par ajout d'attribut. Par exemple, à partir de $Y \cup \{a\}$ et $Y \cup \{b\}$, on produit par fusion $X = Y \cup \{a, b\}$. Il ne reste plus qu'à parcourir le treillis à la recherche des motifs intéressants.

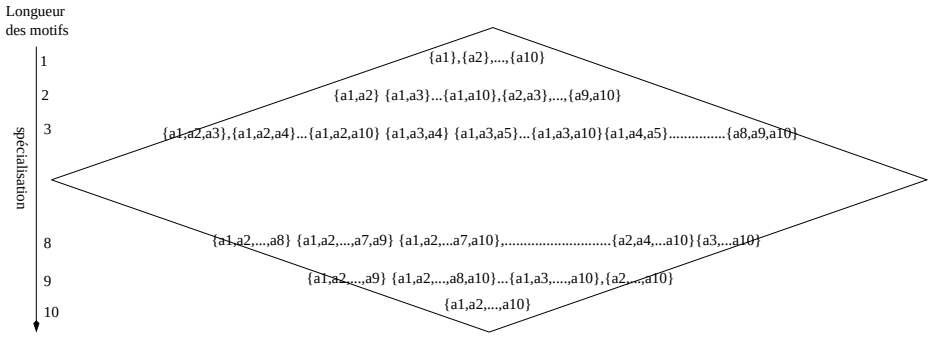

FIG. 1 – Treillis des motifs d'attributs

2.2 Extraction sous contrainte anti-monotone

Nous cherchons à calculer $Th(bd, \mathcal{L}_{\mathcal{A}}, q)$, la théorie ou ensemble des motifs appartenant à $\mathcal{L}_{\mathcal{A}}$ qui satisfont un prédicat q (on écrira indifféremment *contrainte*). Un choix très classique pour q est : le motif est-il *fréquent* dans la base bd ? (son support dépasse-t-il un seuil γ fixé ?), est-il $rare$? (le contraire), mais on pourra également se demander si le motif contient un certain sous-motif, s'il est libre (ie. ne contient pas d'association), fermé, etc.

Certains prédicats ont des propriétés mieux adaptées que d'autres à une recherche dans le treillis, et en particulier ceux qui respectent la relation de spécialisation, ou la relation duale de généralisation. Génératrice de l'espace à parcourir, elle maintient un lien entre les motifs qui guide la recherche du prédicat.

En ce qui concerne le critère de fréquence, la spécialisation d'un motif rare ne peut être que rare. Il s'agit en effet d'une contrainte *monotone*, c'est à dire qu'elle est préservée par la relation de spécialisation. Symétriquement, une contrainte *anti-monotone* (comme le fait d'être fréquent) est préservée par *généralisation* : q_{am} est anti-monotone si $(q_{am}(X) \wedge Y \subset X) \Rightarrow q_{am}(Y)$. Cela signifie que lorsqu'un motif du treillis vérifie q_{am}, alors tous ceux qui sont au dessus également. Ou, quand un motif ne vérifie pas q_{am}, ses spécialisations du dessous non plus.

Les contraintes anti-monotones fournissent ainsi deux critères d'élagage :

Définition 1 (Critère 1)
Si un motif X ne vérifie pas q_{am}, ses spécialisations ne peuvent vérifier q_{am} : il est inutile de les examiner et le treillis peut être élagué sous X.

Définition 2 (Critère 2)
Si le candidat X à la vérification de q_{am} contient un motif qui ne vérifie pas le prédicat, alors ce candidat doit être rejeté : le treillis est élagué au-dessus de X.

Sur la figure 2, les deux générateurs $Y \cup \{a\}$ et $Y \cup \{b\}$ sont représentés, ainsi que leur fusion, $X = Y \cup \{a, b\}$. Si on détecte que $Y \cup \{b\}$ ne vérifie pas la contrainte, alors

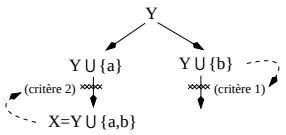

FIG. 2 – Critères d'élagage

la branche qui en est issue est coupée (critère 1). De plus, X contient un motif incorrect $Y \cup \{b\}$: il ne peut donc convenir lui aussi (critère 2).

L'algorithme par niveaux (Mannila & Toivonen, 1997), popularisé par le cas particulier de la contrainte anti-monotone de fréquence, se dessine alors : partant des singletons (niveau 1), on produit chaque motif d'un niveau à l'aide des motifs du niveau précédent, sous réserve qu'il vérifie les deux critères de la contrainte anti-monotone.

2.3 Bordures de théorie

Une utilisation judicieuse de l'anti-monotonie de la contrainte recherchée permet de se contenter des motifs maximaux (au sens de l'inclusion) qui vérifient la théorie. En effet, la relation de généralisation préserve l'anti-monotonie et l'ensemble de la théorie peut alors être générée à partir des motifs maximaux (Mannila & Toivonen, 1997). Cet ensemble des motifs maximaux est appelé la bordure positive de la théorie :

Définition 3 (Bordure positive)
$Bd^+(Th(bd, \mathcal{L}_\mathcal{A}, q)) = \{\varphi \in Th(bd, \mathcal{L}_\mathcal{A}, q) \mid \forall \theta \ \ \varphi \subset \theta \ \Rightarrow \ \theta \notin Th(bd, \mathcal{L}_\mathcal{A}, q)\}$

Le concept dual de la notion de bordure négative est celui de bordure négative :

Définition 4 (Bordure négative)
$Bd^-(Th(bd, \mathcal{L}_\mathcal{A}, q)) = \{\varphi \in \mathcal{L}_\mathcal{A} \backslash Th(bd, \mathcal{L}_\mathcal{A}, q) \mid \forall \theta \ \ \theta \subset \varphi \ \Rightarrow \ \theta \in Th(bd, \mathcal{L}_\mathcal{A}, q)\}$

Un motif appartient à la bordure négative s'il ne vérifie pas la théorie mais tous ses sous-ensembles la vérifient.

Définition 5 (Bordure)
La bordure d'une théorie est la réunion des bordures positive et négative.

$$Bd(Th(bd, \mathcal{L}_\mathcal{A}, q)) = Bd^+(Th(bd, \mathcal{L}_\mathcal{A}, q)) \cup Bd^-(Th(bd, \mathcal{L}_\mathcal{A}, q))$$

Sur la figure 3, le treillis est coupé en deux : en haut la théorie, avec la bordure positive, constituée des motifs maximaux satisfaisant la contrainte, par exemple "de support minimum 3". En bas, les motifs qui n'appartiennent pas à la théorie. Les motifs minimaux de cette partie constituent la bordure négative. La dualité évoquée précédemment souligne le fait que ces bordures sont relatives à une relation de spécialisation et une contrainte anti-monotone. Car sur la figure, la partie basse non hachurée du treillis peut s'interpréter comme la théorie relative à la négation du prédicat (les motifs de support maximum 3), la bordure positive devenant la bordure négative et vice-versa. Ce point sera détaillé à la section 3.1.

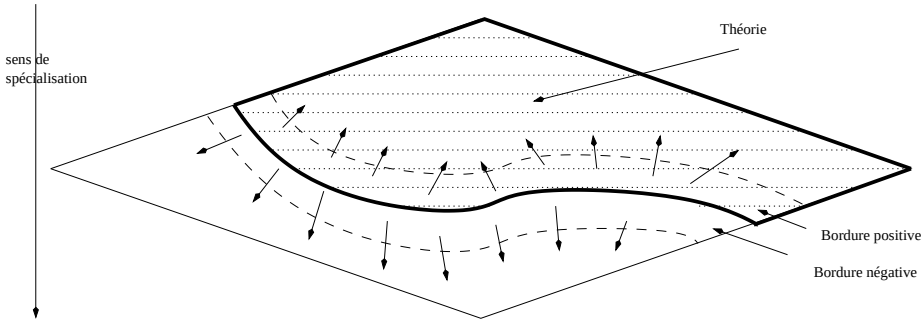

FIG. 3 – Bordures de théorie

2.4 Complexité

Le résultat central de la complexité de l'extraction de théorie indique que l'on ne peut pas faire l'économie de scruter le treillis jusqu'à avoir parcouru la bordure en entier. Un algorithme par niveau effectuera au minimum $|Th(bd, \mathcal{L}_{\mathcal{A}}, q) \cup Bd^{-}(Th(bd, \mathcal{L}_{\mathcal{A}}, q))|$ tests du prédicat sur la base (Mannila & Toivonen, 1997).

Le nombre de tests du prédicat sur la base est crucial, particulièrement dans les gros volumes de données, et la lenteur inhérente au parcours de nombreuses lignes peut s'avérer pénalisante. Énoncée précédemment, la borne inférieure sur les interrogations de la base est donc précieuse : elle mesure les économies d'accès aux données, directement consécutives à l'élagage permis par le critère 1. En revanche, on ne dispose pas de résultat sur l'économie de recherche réalisée par le critère 2. Pourtant, lors de chaque tentative de création d'un candidat, il faudra tester si ses sous-ensembles vérifient la contrainte, ce qui peut s'avérer le goulet d'étranglement de l'algorithme, et ce dès les premiers niveaux.

Dans la suite de cet article, nous reviendrons sur le nombre d'échecs à ce critère 2, particulièrement critique dans le cas de données comportant un grand nombre d'attributs par rapport au nombre d'objets.

2.5 Connexion de Galois

Jusqu'ici, nous avons essentiellement parlé de motifs d'attributs. Or, par le biais de la relation R de la base de données, ces motifs sont connectés à des motifs d'objets. Pour cela, on définit sur $bd = (\mathcal{O}, \mathcal{A}, R)$ les opérateurs f et g de connexion de Galois entre un motif X d'attributs de $\mathcal{A}$ et un motif T d'objets de $\mathcal{O}$:

Définition 6 (Connexion de Galois)
$f(T) = \{a \in \mathcal{A} \mid \forall o \in T, (o, a) \in R\}$ *et* $g(X) = \{o \in \mathcal{O} \mid \forall i \in X, (o, i) \in R\}$. *$f$ représente l'ensemble de tous les attributs communs à un groupe d'objets T (on parle d'intention) et g l'ensemble des objets partageant les mêmes attributs X (extension). Le couple (f, g) définit la connexion de Galois (Birkhoff, 1967) entre $\mathcal{O}$ et $\mathcal{I}$, et $h = f \circ g$ et $h' = g \circ f$ sont les opérateurs de la fermeture de Galois.*

Ainsi, de la même façon qu'on a su représenter le treillis des motifs d'attributs, on peut représenter le treillis des motifs d'objets. Les motifs des deux treillis sont connectés par les opérateurs f et g. Les concepts symbolisent la passerelle et associent deux motifs fermés :

Définition 7 (Motif fermé)
Un motif X d'attributs est fermé ssi $h(X) = X$. Un motif T d'objets est fermé ssi $h'(T) = T$. Un concept (X, T) associe deux fermés X d'attributs et T d'objets, tels que $X = f(T)$ (ou $T = g(X)$).

Dans le contexte de la fouille de données, les motifs fermés sont bien connus, entre autres parce que leurs propriétés sont multiples : ils permettent de calculer efficacement les supports (Bastide *et al.*, 2002), déterminer les ensembles minimaux de règles d'association (Pasquier *et al.*, 1999; Zaki, 2000; Luong, 2001), favorisent le clustering (Durand & Crémilleux, 2002), etc. Nous les utilisons ici pour la passerelle qu'ils définissent entre objets et attributs. Dans les faits, on constate que les opérateurs f et g conservent la propriété de fermeture (si X est fermé, $g(X)$ l'est également, de même pour T et $f(T)$). Ces propriétés d'invariance par les opérateurs de fermeture fournissent un lien bidirectionnel fort entre un motif d'attributs et un motif d'objets : la connexion de Galois.

De plus, cette invariance assure qu'un motif fermé sera connecté à un autre motif fermé. Connaissant un fermé d'attributs, il est possible de passer à un fermé d'objets et vice-versa. Cette passerelle sera utilisée Section 4, appliquée sur la collection complète des fermés d'un type pour déduire les fermés de l'autre type mis en jeu.

La notion de théorie relative à un prédicat dans une base de données peut alors être restreinte des motifs d'attributs aux concepts :

Définition 8 (Théorie des concepts)
La théorie des concepts relativement à une base de données bd, le langage $\mathcal{L}$ associé, et une contrainte q, notée $Th_{\mathcal{C}}(bd, \mathcal{L}, q)$, est l'ensemble des concepts (X, T) tels que X appartient à $Th(bd, \mathcal{L}_{\mathcal{A}}, q)$.

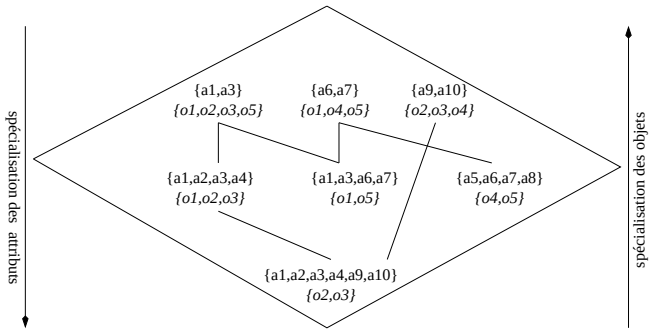

FIG. 4 – Treillis de Galois

Sur la figure 4, nous avons représenté le même treillis qu'à la figure 1, restreint aux motifs fermés. Ce treillis est usuellement appelé treillis de concepts (Simon, 2000; Wille, 1992). Chaque motif est un concept, étiqueté à la fois par une liste d'attributs et une liste d'objets. On remarque que la relation de spécialisation des attributs, qui oriente le treillis vers le bas, est désormais accompagnée d'une même relation de spécialisation, sur les objets cette fois, mais en sens inverse. En effet, la connexion inverse le sens d'inclusion : si $X \subset Y$ alors $g(X) \supseteq g(Y)$.

3 Nouvelle méthode d'extraction par transposition

Nous proposons dans cette section une nouvelle méthode d'extraction, qui tire parti de la connexion de Galois entre les motifs d'attributs et d'objets. Nous donnons les définitions de la transposition de base de données, puis de la transposition de prédicat et finissons par énoncer le résultat central de complémentarité de l'extraction des motifs et de sa transposée, que nous avons mis en exergue pour proposer notre méthode.

3.1 Transposition de base de données

Muni de la connexion de Galois, le double treillis représente les motifs à la fois d'attributs et d'objets, avec chacun son sens particulier de spécialisation (cf. Figure 4). Il est donc possible d'y réaliser deux extractions :

- sur les attributs, en partant du haut et suivant la relation de spécialisation côté attribut,
- sur les objets, en partant du bas et suivant la relation de spécialisation côté objet.

Si $bd = (\mathcal{O}, \mathcal{A}, R)$, la première extraction calcule $Th(bd, \mathcal{L}_\mathcal{A}, q)$ où $\mathcal{L}_\mathcal{A}$ est le langage des motifs d'attributs. La deuxième extraction calcule la théorie relative à la *transposée* de bd sur le langage des motifs d'objets :

Définition 9 (Base transposée)
Soit $bd = (\mathcal{O}, \mathcal{A}, R)$ un base de données. La base transposée s'écrit ${}^t bd = (\mathcal{A}, \mathcal{O}, {}^t R)$ où $(a, o) \in {}^t R \iff (o, a) \in R$.

3.2 Transposition de prédicat

Si la transposition de base de données est une chose relativement naturelle, il n'en est pas de même pour le prédicat qui contraint la recherche. Dans le cas de la contrainte de fréquence, la notion duale de support pour les motifs d'attributs est la longueur du motif d'objets correspondant. Par exemple, sur la figure 4, le motif $\{a_9, a_{10}\}$ est présent dans les objets o_2, o_3, o_4 : son support est 3, la longueur de $\{o_2, o_3, o_4\}$. Nous précisons cette notion en utilisant la connexion de Galois pour passer du formalisme des attributs aux objets :

Définition 10 (Prédicat transposé)
Soit q un prédicat sur le langage $\mathcal{L}_\mathcal{A}$. Son transposé ${}^t q$ est défini sur $\mathcal{L}_\mathcal{O}$ par (f est

l'opérateur de Galois) :

$$\forall T \in \mathcal{L}_{\mathcal{O}},\ {}^{t}q(T) \iff q(f(T))$$

Par exemple, si q indique que le motif d'attributs a une fréquence supérieure au seuil γ, le prédicat transposé ${}^{t}q$ indiquera que le motif d'objets a une longueur supérieure à γ.

Relativement à la spécialisation des attributs, ${}^{t}q$ sera monotone (resp. anti-monotone) si q est monotone (resp. anti-monotone) également. Or, la spécialisation sur les objets suit le sens inverse des attributs ; si q est anti-monotone suivant les attributs, ${}^{t}q$ est monotone suivant la spécialisation des objets et il faut en prendre la négation pour retrouver le prédicat anti-monotone qui peut guider notre recherche. Nous obtenons la propriété suivante :

Propriété 1
Si q est anti-monotone vis à vis de la spécialisation des attributs, alors $\neg {}^{t}q$ est également anti-monotone vis à vis de la spécialisation des objets.

Preuve : La connexion de Galois inverse le sens d'inclusion : q, anti-monotone suivant les attributs fournit un prédicat transposé ${}^{t}q$ qui est monotone suivant les objets : sa négation est anti-monotone.

Pour illustrer cette propriété, revenons à la contrainte de fréquence sur les motifs d'attributs. On s'intéressera par exemple aux motifs présents au moins trois fois dans la base. La transposition est une contrainte sur la longueur des motifs d'objets, qui requiert donc des motifs de longueur supérieure à trois. Cette nouvelle contrainte est monotone relativement à la spécialisation des objets (les spécialisations d'un motif de plus de trois objets contiennent également plus de trois objets). La négation de la contrainte transposée, ie. exiger que les motifs contiennent moins de trois objets, est donc anti-monotone.

3.3 Extraction transposée

Nous disposons désormais d'une opération de transposition de base de données et d'un nouveau prédicat anti-monotone relativement à la spécialisation des objets : toutes les conditions sont réunies pour pouvoir appliquer l'algorithme par niveaux classique. Néanmoins, obtenir le nouveau prédicat nécessite de transposer l'ancien mais surtout de le nier afin de garantir son anti-monotonie. La nouvelle extraction produira alors le complémentaire de la théorie fournie par l'extraction originale.

Définition 11 (Théorie des concepts transposée)
Sur ${}^{t}bd$ avec la contrainte $\neg {}^{t}q$, l'algorithme par niveaux extraira la théorie des concepts $Th_{\mathcal{C}}({}^{t}bd, \mathcal{L}, \neg {}^{t}q)$, transposée de $Th_{\mathcal{C}}(bd, \mathcal{L}, q)$.

La propriété suit immédiatement :

Propriété 2 (Complémentarité des extractions)
Relativement à l'intégralité des concepts, la théorie des concepts $Th_{\mathcal{C}}(bd, \mathcal{L}, q)$ et sa transposée $Th_{\mathcal{C}}({}^{t}bd, \mathcal{L}, \neg {}^{t}q)$ sont complémentaires.

Preuve : Définis sur les attributs ou les motifs, q et son transposé ${}^t q$ sont équivalents (leurs théories contiennent les mêmes concepts). Les théories relatives à q et ${}^t q$ sont donc identiques : relativement à q et $\neg {}^t q$ elles sont complémentaires.

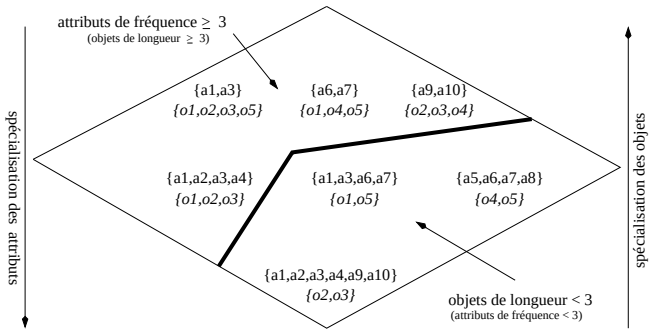

FIG. 5 – Complémentarité des extractions

Sur notre exemple jouet, les motifs fermés de support minimum 3 sont $\{a_1, a_3\}$, $\{a_1, a_2, a_3, a_4\}$, $\{a_6, a_7\}$ et $\{a_9, a_{10}\}$. En transposant puis niant cette contrainte, on recherchera donc les motifs d'objets de longueur maximale 3 : on trouve les motifs $\{o_1, o_5\}$, $\{o_4, o_5\}$ et $\{o_2, o_3\}$. On vérifie alors que les deux ensembles de concepts sont complémentaires (voir figure 5).

Cette propriété de complémentarité justifie notre nouvelle méthode d'extraction en travaillant sur la base transposée, car il est possible, en transposant, d'obtenir le complémentaire de la théorie initiale recherchée, et donc facilement la théorie elle-même. On peut aussi tirer parti de la dualité des bordures exposée à la fin de la section 2, car la bordure de la théorie de la base originale coïncide, grâce à la complémentarité, avec la bordure de la théorie issue de la transposée. Il s'agit de mettre à profit les algorithmes efficaces d'extraction directe de la bordure (Mitchell, 1980; Gunopulos *et al.*, 1997; Zaki *et al.*, 1997; Bayardo, 1998), et la bordure négative dans la base transposée fournit directement et sans aucun sur-coût la bordure positive dans la base originale. Les différents aspects de cette méthode sont discutés dans la section suivante.

4 Applications

Dans cette section, nous détaillons les cas d'utilisation de l'extraction sur la transposée, et présentons plusieurs exemples d'application. La première opération intéressante consiste à extraire complètement les motifs fermés d'objets pour déduire, moyennant un coût nul, des informations sur les fermés d'attributs. Intéressante pour les situations où l'extraction normale échoue du fait des dimensions de la base, cette approche est validée sur des données biologiques. Ensuite, nous montrons les améliorations de performances relativement au critère 2 d'élagage, et nous terminons par une méthode d'extraction sous contrainte monotone.

	jouet		^{t}jouet	
long.	succès	échecs	succès	échecs
1	32	13	9	1
2	24	22	4	4
Total	56	35	13	5
libres	38		14	
fermés	10			

TAB. 2 – Échecs/succès du critère d'élagage sur base `jouet`

4.1 Extraction de l'ensemble des motifs fermés

Un algorithme calculant l'ensemble des motifs fermés sur la base de données les fournit sous forme d'attributs, et la connexion de Galois déduit les fermés d'objets correspondants. Réciproquement, l'extraction transposée fournit les fermés d'objets, que la connexion de Galois traduit en fermés d'attributs.

Il est donc possible d'obtenir le même ensemble de concepts en extrayant depuis la base ou sa transposée. Le choix de l'une ou l'autre méthode sera guidé par les dimensions de la base : entre le nombre d'attributs ou d'objets, on choisit le plus petit, celui qui génèrera le treillis le moins vaste.

Sur notre exemple jouet, la plus petite dimension est celle des objets : 5 motifs qui donnent lieu à $2^5 = 32$ motifs. Parmi ces 32 motifs, seuls 10 sont fermés. En revanche, sous sa forme originale, la base contient 10 attributs soit $2^{10} = 1024$ motifs, mais il n'y en a toujours que 10 qui sont fermés. Il est clair qu'il est plus efficace de les extraire en choisissant la dimension la plus faible. Ces données sont reprises dans le tableau récapitulatif 2.

Différentes techniques existent pour obtenir l'ensemble des fermés. Nous utilisons personnellement les motifs *libres* (Boulicaut *et al.*, 2000) ou motifs *clés* (Bastide *et al.*, 2002) pour générer les fermés (un motif libre est un motif pour lequel il n'y a pas de règle d'association, cf. (Boulicaut & Bykowski, 2000)). De plus, la contrainte de liberté est anti-monotone, ce qui n'est pas le cas de celle de fermeture, et fournit donc un critère d'élagage pour l'algorithme par niveaux. En revanche, un même fermé peut être généré par plusieurs libres. Sur notre exemple, l'extraction des libres fournit 38 motifs libres qui génèrent les 10 fermés, alors que l'extraction dans la transposée se contente de 14 libres. Le gain est immédiat.

Nous avons également appliqué notre méthode à des données intitulées `sain8`, provenant d'une expérience réalisée à l'unité INRA/INSERM U449 (Rome *et al.*, 2003). Il s'agit de l'analyse du transcriptome de biopsies musculaires humaines, avant et après trois heures de clamp euglycémique-hyperinsulinémique. La matrice d'expression résultante contient 6 lignes pour 1065 colonnes. Dans cette disposition des données, l'extraction fournit en quelques minutes 667 831 motifs libres pour seulement 41 fermés. Dans la matrice transposée, l'extraction ne prend que quelques centièmes de secondes et se contente de 42 motifs libres, pour les mêmes 41 fermés.

	sain8		^{t}sain8	
long.	succès	échecs	succès	échecs
1	777	0	6	0
2	172 548	128 928	15	0
3	2 315 383	4 713 114	16	4
4	2 965 726	9 371 325	6	9
5	0	1 544 485	0	2
Total	5 454 434	15 757 852	43	15
libres	667 831		42	
fermés	41			

TAB. 3 – Échecs/succès du critère d'élagage sur base sain8

Cet exemple est tout à fait symptomatique de l'efficacité de l'extraction dans la transposée. Pour les contextes biologiques et leurs bases de données de dimensions pathologiques, cette amélioration est spectaculaire.

D'une manière plus générale, l'utilisation de la tranposition pour obtenir à moindre frais la totalité des fermés est justifiée dans tous les contextes où le nombre d'exemples est faible devant le nombre de descripteurs. Cette situation extrême interdit souvent l'usage des algorithmes classiques, historiquement conçus pour traiter des bases comportant de nombreuses lignes. Mais en médecine et en biologie, les expériences sont coûteuses et il faut se contenter d'un petit nombre de situations. Remarquons que même si la fouille de texte est mieux dotée en exemples, celle-ci peut être gênée par de trop nombreux attributs. L'extraction des concepts dans la transposée offre alors une opportunité de pallier un problème classique de dimensions. Enfin, en tant que représentation condensée des motifs, la collection des concepts autorise de multiples usages : calcul des supports, règles d'association, règles de classification, clustering, etc.

4.2 Critère d'élagage

Dans le tableau 3, nous présentons pour sain8 et sa transposée, le nombre de motifs par niveau de l'algorithme qui ont réussi les deux critères et qui devront être examinés dans la base, et le nombre d'échecs du critère 2. À chaque niveau, les candidats sont générés. Ceux qui passent le test avec succès rejoindront les motifs libres après confrontation à la base, les autres conduisent à un échec pour absence de l'un de ses sous-motifs.

L'extraction dans ^{t}sain8 est exceptionnellement efficace : elle produit moins de motifs à tester dans la base (37 contre 5 453 657) et elle conduit à en refuser infiniment moins : 13 contre 14 213 367.

Dans les contextes biologiques, cet argument est particulièrement pertinent. Une extraction sur peu de lignes mais de nombreuses colonnes échouera par manque de ressource mémoire ou à cause d'un temps d'exécution rédhibitoire. Même s'il y a peu de lignes et si les passes sur la base ne pénalisent pas l'algorithme, le nombre d'échecs au critère 2 dégrade les performances et de trop nombreux candidats sont générés, pour un

rendement très faible.

Somme toute, le résultat sur la complexité des algorithmes d'extraction précise que l'on examinera au moins autant de motifs qu'il y en a dans la bordure. Or nous avons indiqué que les bordures étaient symétriquement communes aux deux extractions sur la base et sa transposée. Si cette borne inférieure ne peut donc être améliorée, la différence se jouera sur le nombre de motifs que l'on évitera de générer puis d'inspecter.

4.3 Extraction sous contrainte monotone

Les contraintes anti-monotones fournissent une classe importante de contrainte qui rendent l'extraction possible. Mais il en existe bien d'autres également très utiles, et bien sûr les contraintes monotones ! Par exemple, chercher l'ensemble des motifs rares dans une base (Li *et al.*, 1999; Dong & Li, 1999; L. De Raedt, 2001) utilise une contrainte monotone. L'extraction sous contrainte quelconque est un domaine de recherche à part entière (Jeudy, 2002) et des algorithmes extrayant simultanément les deux types existent (Bucila *et al.*, 2002). Une démarche consiste à morceler la contrainte en conjonction et disjonction de contraintes monotones, anti-monotones. Nous devons donc disposer de méthodes capables de traiter les deux types basiques, monotone et anti-monotone.

Le caractère monotone de q interdit les algorithmes classiques, fondés sur l'exploitation de contraintes anti-monotone qui servent à élaguer l'espace de recherche. Mais d'après la propriété de complémentarité, $Th(bd, \mathcal{L}, q)$ et $Th({}^t bd, \mathcal{L}, \neg {}^t q)$ fournissent la théorie pour q et son complémentaire. D'après la propriété 1, $\neg {}^t q$ est également monotone, donc $Th({}^t bd, \mathcal{L}, \neg {}^t q)$ n'est pas non plus calculable simplement. En revanche, ${}^t q$ est anti-monotone pour la relation de spécialisation des objets, et ce qui précéde prouve ainsi la propriété suivante :

Propriété 3
$Th_{\mathcal{C}}(bd, \mathcal{L}, q)$, la théorie des concepts relative à une propriété monotone q est :

$$Th_{\mathcal{C}}({}^t bd, \mathcal{L}, {}^t q)$$

Suivant une contrainte monotone, l'extraction des motifs peut donc être menée en utilisant simplement l'algorithme par niveaux, avec la contrainte transposée des attributs aux objets, dans la base transposée. Par exemple, si on cherche les motifs rares, dont le support est inférieur à 3, c'est bien une contrainte monotone. La transposition de la contrainte requiert que les motifs d'objets soient de longueur inférieure à 3. C'est bien une contrainte anti-monotone relativement à la spécialisation des objets. L'extraction sur la transposée de la base de données est possible avec l'algorithme par niveaux et fournit alors les concepts recherchés.

5 Conclusion - Perspectives

Nous avons expliqué que l'extraction de théorie relative à une contrainte dans une base de données est une tâche ardue dans les très larges volumes de données (de

nombreux attributs), un cas classique dans les données génomiques. Nous avons alors proposé une nouvelle méthode d'extraction, fondée sur l'exploitation des transpositions de la base de données et de la contrainte. Puis nous avons montré l'utilité de cette technique pour obtenir plus facilement l'intégralité des fermés, et économiser sur le coût de l'algorithme en minimisant l'espace à parcourir et les échecs des critères d'élagage. Enfin nous avons proposé un nouveau procédé d'extraction sous contrainte monotone.

Les perspectives d'application sont nombreuses, précisément dans les domaines d'étude du génome. Cependant, il reste à compléter les outils qui peuvent exploiter cette méthode. En particulier, des algorithmes d'extraction en profondeur qui fournissent directement la bordure commune peuvent être adaptés pour tirer parti de cette propriété.

Enfin, il sera très utile d'étudier les transpositions de contraintes. Si le passage de la contrainte de support sur les attributs se fait simplement à la longueur des objets, il n'en va pas aussi naturellement de la contrainte de liberté (pas d'association présente dans le motif) par exemple.

Remerciements. Les auteurs remercient Sophie Rome et Hubert Vidal (INRA-INSERM U449) pour la fourniture des données et Jeremy Besson (LIRIS CNRS FRE 2672) pour le travail de pré-traitement. François Rioult est financé par l'Unité IRM du CHU de Caen, le Comité de la Manche de la Ligue contre le Cancer et le Conseil Régional de Basse-Normandie.

Références

AGRAWAL R., MANNILA H., SRIKANT R., TOIVONEN H. & INKERI VERKAMO A. (1996). Fast discovery of association rules. *Advances in Knowledge Discovery and Data Mining*.

BASTIDE Y., TAOUIL R., PASQUIER N., STUMME G. & LAKHAL L. (2002). Pascal : un algorithme d'extraction des motifs fréquents. *Technique et science informatiques. Vol. 21 - n° 1*, p. 65–95.

BAYARDO R. (1998). Efficiently mining long patterns from databases. *ACM-SIGMOD*, p. 85–93.

BIRKHOFF G. (1967). Lattices theory. *American Mathematical Society, vol. 25*.

BOULICAUT J.-F. & BYKOWSKI A. (2000). Frequent closures as a concise representation for binary data mining. *PAKDD 2000*.

BOULICAUT J.-F., BYKOWSKI A. & RIGOTTI C. (2000). Approximation of frequency queries by means of free-sets. In *Principles of Data Mining and Knowledge Discovery (PKDD'00)*, p. 75–85.

BUCILA C., GEHRKE J., KIFER D. & WHITE W. (2002). Dualminer : A dual-pruning algorithm for itemsets with constraints. *Proceedings of SIGKDD'02*.

DONG G. & LI J. (1999). Efficient mining of emerging patterns : discovering trends and differences. *Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining, San Diego, USA (SIGKDD'99)*, p. 43–52.

DURAND N. & CRÉMILLEUX B. (2002). Ecclat : a new approach of clusters discovery in categorical data. *22nd Int. Conf. on Knowledge Based Systems and Applied Artificial Intelligence (ES'02)*, p. 177–190.

GUNOPULOS D., MANNILA H., KHARDON R. & TOIVONEN H. (1997). Data mining, hypergraph transversals, and machine learning. In *PODS 1997*, p. 209–216.

JEUDY B. (2002). Optimisation de requêtes inductives : application à l'extraction sous contrainte de règles d'association. *Ph.D. Thesis at INSA of Lyon*.

L. DE RAEDT S. K. (2001). The levelwise version space algorithm and its application to molecular fragment finding. *In proceedings of IJCAI'01*, p. 853–862.

LI J., ZHANG X., DONG G., RAMAMOHANARAO K. & SUN Q. (1999). Efficient mining of high confidence association rules without support thresholds. *Proceedings of PKDD 99 – 3rd European Conference on Principles and Practice of Knowledge Discovery in Databases*.

LUONG V. P. (2001). Reasoning on association rules. *BDA'2001, 17è Journées Bases de Données Avancées*.

MANNILA H. & TOIVONEN H. (1996). Multiple uses of frequent sets and condensed representations (extended abstract). In *Knowledge Discovery and Data Mining*, p. 189–194.

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

MITCHELL T. (1980). Generalization as search. *Artificial Intelligence, vol. 18, n 2 p. 203-226*.

PASQUIER N., BASTIDE Y., TAOUIL R. & LAKHAL L. (1999). Closed set based discovery of small covers for association rules. In *Proc. 15emes Journees Bases de Donnees Avancees, BDA*, p. 361–381.

ROME S., CLÉMENT K., RABASA-LHORET R., LOIZON E., POITOU C., BARSH G. S., RIOU J.-P., LAVILLE M. & VIDAL H. (2003). Microarray profiling of human skeletal muscle reveals that insulin regulates 800 genes during an hyperinsulinemic clamp. *Journal of Biological Chemistry*.

SIMON A. (2000). Outils classificatoires par objets pour l'extraction de connaissance dans les bases de données. *Doctorat de l'Université Henri-Poincaré, Nancy 1*.

WILLE R. (1992). Concept lattices and conceptual knowldge systems. *Computer mathematic applied, 23(6-9) :493-515*.

ZAKI M., PARTHASARATHY S., OGIHARA M. & LI W. (1997). New algorithms for fast discovery of association rules. *In 3rd Intl. Conf. on Knowledge Discovery and Data Mining*, p. 283–296.

ZAKI M. J. (2000). Generating non-redundant association rules. *6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Boston*, p. 34–43.