Régler le processus d'exploration dans l'analyse relationnelle de concepts - Le cas de données hydroécologiques

Amirouche Labib Ouzerdine, Agnès Braud, Xavier Dolques, Florence Le Ber, Marianne Huchard

Régler le processus d'exploration dans l'analyse relationnelle de concepts – Le cas de données hydroécologiques

Amirouche Ouzerdine*, Agnès Braud**, Xavier Dolques***,
Marianne Huchard*, Florence Le Ber**

* LIRMM, Univ Montpellier, CNRS, Montpellier, France
labib23dz@hotmail.com, huchard@lirmm.fr
** ICube, Université de Strasbourg, CNRS, ENGEES, France
agnes.braud@unistra.fr, florence.leber@engees.unistra.fr,
*** Movidone, Strasbourg, France
xavier.dolques@laposte.net

Résumé. Cet article s'intéresse à l'exploration de jeux de données multi-relationnelles, et aux différentes manières de les analyser en utilisant l'analyse relationnelle de concepts (ARC), une variante de l'analyse formelle de concepts. L'ARC utilise plusieurs quantifieurs d'échelonnage qui rendent le processus d'analyse finement réglable, permettant une grande flexibilité dans l'exploration et dans ses résultats. En contrepartie, l'analyste peut être submergé par l'ensemble des choix qu'il doit faire au cours de l'analyse. Pour traiter ce problème, nous proposons trois sur-couches qui aident l'analyste à anticiper et contrôler les résultats de ses choix. Notre proposition est appliquée à un jeu de données sur la qualité des eaux de rivières.

1 Introduction

Les jeux de données multi-relationnelles suivent un schéma (modèle de données), où des entités (objets) de plusieurs catégories sont décrites par leurs caractéristiques (attributs) et où des relations relient les objets des différentes categories. Les experts des domaines concernés les exploitent dans le cadre de différentes tâches : consultation ou exploration, requêtage, extraction de motifs, ou encore classification, au sens d'organiser dans une hiérarchie de généralisation des ensembles d'objets similaires. Cet article se concentre sur la tâche d'*exploration* fondée sur une telle organisation hiérarchique.

L'analyse formelle de concepts, ou AFC (Ganter et Wille, 1999), et ses extensions apportent des méthodes qui contribuent à l'exploitation des données, notamment multi-relationnelles (Liquière et Sallantin, 1998; Kötters, 2013; Ferré, 2015; Ferré et al., 2005; Hacene et al., 2013). Parmi ces méthodes, l'analyse relationnelle de concepts (ARC) a été spécialement conçue pour les tâches sus-citées (Hacene et al., 2013). Cette méthode construit un ensemble de classsifications interconnectées, qui peuvent être utilisées pour extraire des motifs et des règles d'implication portant sur les liens inter-objets ainsi que sur des abstractions de ces liens.

L'ARC a pour caractéristique principale de construire des abstractions des liens inter-objets en appliquant des opérations d'échelonnage sur des groupes d'objets. Ces abstractions per-

mettent de grouper des objets qui ont *au moins un / seulement un / tous / au moins 30% / etc.* (parmi) leurs liens sortants pour une certaine relation vers un autre groupe d'objets identifié ; ces quantifieurs sont appelés quantifieurs d'échelonnage. Ceci permet de propager des regroupements d'objets à travers des chaînes de liens inter-objets. Combinée avec la nature incrémentale et exploratoire de l'ARC, où à chaque étape on peut choisir les relations à considérer, la variété des quantifieurs rend le processus très finement réglable (Braud et al., 2018).

D'un côté, ces possibilités de réglage rendent les résultats de l'ARC très expressifs, avec des descriptions utilisant des quantifieurs autres que les classiques quantifieurs universel et existentiel. D'un autre côté, la multiplicité des choix à réaliser lors d'une tâche d'analyse peut submerger l'analyste. Pour traiter ce problème, une adaptation de l'ARC a été proposée par Dolques et al. (2015), où les relations sont explorées de façon graduelle grâce à une configuration pas-à-pas plutôt que globale. Néanmoins, à notre connaissance, aucun outil n'existe qui permette de guider l'utilisateur et de rendre le processus plus intuitif.

Dans cet article, nous proposons d'introduire de la connaissance dans le processus de l'ARC pour faciliter l'analyse : de la connaissance a priori fournie par l'utilisateur pour contraindre le processus au départ, et de la connaissance fournie par le processus à l'utilisateur, qui peut ainsi mieux décider comment le régler. Nous proposons trois sur-couches du processus de l'ARC : la première permet d'exprimer des contraintes sur les quantifieurs, pour en faire un choix cohérent ; la seconde permet de traduire des requêtes de haut niveau (du niveau schéma), qui sont difficiles à exprimer par les utilisateurs, en des expressions d'un langage contrôlé ; la troisième permet de donner des métriques quantitatives sur les treillis et sur les règles d'implication obtenus par des réglages voisins du réglage courant afin d'aider les utilisateurs à affiner leur analyse. Nous appliquons nos propositions à un jeu de données sur la qualité des eaux de rivières, issu du projet FRESQUEAU [1], qui concerne les relations entre l'état physico-chimique de l'eau et les caractéristiques des taxons (macroinvertebrés) qui y vivent. Ce jeu de données est réduit à des fins d'illustration : des jeux de données plus importants, et dans d'autres domaines, peuvent être traités par des techniques de seuillage ou de sélection de concepts,

L'article est organisé de la manière suivante. La section 2 présente le type d'exploration relationnelle que l'ARC permet et pourquoi il peut être difficile pour les experts de conduire une analyse. La section 3 présente l'ARC et son utilisation comme outil de requête. La section 4 décrit les trois sur-couches et les illustre sur les données FRESQUEAU. Nous concluons et dressons quelques perspectives dans la section 5.

2 Exploration de jeux de données relationnelles

De nombreuses données sont multi-relationnelles de façon inhérente, impliquant des relations de différentes natures. Ceci motive le développement de méthodes pour extraire des motifs relationnels ou des règles d'association relationnelles, induire des arbres de décision relationnels, ou encore construire des classes en utilisant des distances relationnelles (Džeroski, 2003). Dans le domaine de l'AFC, les données relationnelles sont envisagées par le biais de représentations à base de graphes (Liquière et Sallantin, 1998; Kötters, 2013; Ferré, 2015), sous forme d'expressions logiques (Ferré et al., 2005) ou sous forme tabulaire, comme dans l'ARC (Hacene et al., 2013).

1. http ://engees-fresqueau.unistra.fr

Quand les experts ont une connaissance vague des données et quand leurs questions sont générales, une approche exploratoire s'avère adaptée (Wildemuth et Freund, 2012; Palagi et al., 2017). L'exploration des données peut alors être complètement libre, ou guidée par ces questions générales, souvent au niveau du schéma des données (concepts et relations).

Nous utilisons ici comme illustration un cas d'exploration typique pour un hydroécologue étudiant l'effet de l'état physico-chimique des eaux d'une rivière sur les caractéristiques (ou traits de vie) des taxons qui y vivent. Un extrait du modèle de données est montré sur la figure 1 : les sites échantillonnés (par différentes techniques de prélèvement) ont une certaine abondance de taxons (identifiants d'êtres vivants, ici des macroinvertebrés, organisés en genres et familles) qui ont une certaine affinité pour certaines modalités de traits de vie (p. ex. modalités de taille maximale, stade aquatique – œuf, larve, nymphe – mode de respiration, mode de locomotion). Les sites échantillonnés sont aussi décrits par des mesures sur des paramètres physico-chimiques (PC) (p. ex. nitrites, minéraux, matière organique, température) organisés en catégories selon leur nature. Pour chaque relation, le niveau varie en 5 degrés d'intensité dépendant du nombre d'individus d'un taxon sur le site pour l'*abondance*, de la part de population d'un taxon montrant une modalité d'un trait de vie pour l'*affinité* et de la valeur mesurée sur le site échantillonné pour une *mesure de paramètre PC*. Pour les besoins de l'analyse, chaque relation entre les catégories d'objets est divisée en 5 relations correspondant aux niveaux.

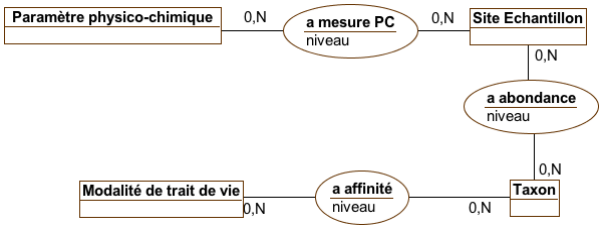

FIG. 1 – *Extrait du modèle de données du projet* FRESQUEAU

Une question générale posée par les hydroécologues est : *quels sont les liens entre les traits de vie des taxons et les valeurs des paramètres physico-chimiques ?* L'exploration du jeu de données pour répondre à cette question peut prendre plusieurs formes, comme l'extraction de règles impliquant les relations, ou le regroupement d'objets de différentes catégories (comme les sites échantillonnés ou les taxons) selon leurs attributs et les objets d'autres catégories avec lesquels ils sont connectés. Par exemple, les experts peuvent être intéressés par les réponses à la question suivante : *trouver des groupes de sites échantillonnés qui ont (1) un certain niveau d'abondance pour un groupe de taxons, ayant eux-même en commun un groupe de traits de vie avec un certain niveau d'affinité et qui ont (2) à un certain niveau, des paramètres physico-chimiques d'un certain groupe*. Les résultats peuvent révéler, par exemple : un groupe de sites contenant des taxons avec une durée de vie longue, et contenant beaucoup de matières organiques ; un groupe de sites avec un haut niveau de minéraux dissous, et contenant des taxons qui se déplacent en rampant (mode de locomotion). Pour une étude plus précise, la question générale et les groupes extraits peuvent être raffinés dans plusieurs directions. Pour en donner seulement un exemple, considérons la question suivante (les termes en caractères gras indiquent des points de variabilité dans la question) :

— *trouver des groupes de sites échantillonnés qui ont (1) un **certain** niveau d'abondance pour un groupe de taxons (et **plus de 60%** des taxons de chaque site sont dans ce groupe), ayant eux-même **seulement** des traits de vie dans un même groupe de traits de vie avec un certain niveau d'affinité et qui ont (2) à un certain niveau de concentration, **tous les** paramètres physico-chimiques d'un certain groupe*.

D'une part, reformuler la question générale avec ces indications est important pour les experts car ils ont besoin de réponses précises et peuvent aussi vouloir changer le focus de leur analyse. D'autre part, le raffinement peut être fait dans de nombreuses directions, et les experts peuvent être rapidement perdus. Par ailleurs, choisir un raffinement plutôt qu'un autre peut conduire à des ensembles de résultats trop restrictifs, ou au contraire, à des ensembles trop larges. Finalement, les résultats sont des ensembles de groupes d'objets reliés qui respectent un certain schéma de question. Ces groupes peuvent être nombreux et peuvent se spécialiser les uns les autres, comme nous le montrons plus loin.

3 L'ARC, un outil pour l'interrogation de données

L'analyse relationnelle de concepts (Hacene et al., 2013) étend l'analyse formelle de concepts (Ganter et Wille, 1999) aux jeux de données multi-relationnelles et ajoute aux approches citées précédemment une panoplie d'opérateurs et une approche itérative qui permet de suivre la propagation de la connaissance, de faciliter la compréhension du processus de formation des motifs, des règles ou des classifications et de favoriser le raisonnement, y compris abductif. Les jeux de données multi-relationnelles y sont représentés par des familles relationnelles de contextes (RCF), composées de contextes formels et de relations. Chaque contexte formel (ou *contexte objets-attributs*) représente un ensemble d'objets d'une certaine catégorie par ses attributs. Une relation (ou *contexte objets-objets*) connecte les objets des différentes catégories (ou de la même catégorie).

Définition 1 (Famille relationnelle de contextes (RCF)) *Une famille relationnelle de contextes est une paire* $(\mathbf{K}, \mathbf{R})$ *où :* $\mathbf{K} = \{\mathcal{K}_i\}_{i=1,\ldots,n}$ *est un ensemble de* $\mathcal{K}_i = (G_i, M_i, I_i)$ *contextes objets-attributs et* $\mathbf{R} = \{r_j\}_{j=1,\ldots,p}$ *est un ensemble de* r_j *relations (contextes objets-objets) où* $r_j \subseteq G_k \times G_l$ *pour* $k, l \in \{1, \ldots, n\}$.

Le tableau 3 montre un exemple simplifié de RCF inspiré de notre application dans le domaine hydro-écologique (a_abondance est ici une relation binaire pour simplifier l'illustration). Le contexte formel `Taxon` introduit :

— les taxons *Aeschnidae* (`Aes.`), *Agabus* (`Agb.`), *Agraylea* (`Aga.`), *Agriotypus* (`Agi.`), *Ancylus* (`Anc.`), *Anisus* (`Ani.`), *Anodonta* (`Ano.`), *Anthomyiidae* (`Ant.`).

— cinq attributs décrivant leur micro-habitat (rochers, graviers, sable, macrophytes, débris organiques).

Le contexte formel `SiteEchantillon` décrit 8 sites par les caractéristiques de leur flux (torrent, eau calme) et leurs composants chimiques. Le contexte objets-objets `a_abondance` connecte les 8 sites aux taxons qui y ont été trouvés durant l'échantillonage.

L'AFC peut être appliquée aux contextes formels `Taxon` et `SiteEchantillon` pour former des hiérarchies de groupes d'objets partageant des attributs communs (concepts).

Taxon	rochers	graviers	sable	macrophytes	debrisOrg
Aes					×
Agb				×	×
Aga				×	×
Agi	×	×	×		×
Anc	×	×			
Ani	×	×	×		×
Ano				×	×
Ant	×	×			

SiteEchantillon	torrent	eauCalme	NH4	SO4	Ca	Mg	C3H8NO5P
se1	×		×				
se2	×		×				
se3		×		×			
se4		×		×			
se5		×		×			×
se6		×		×			×
se7		×		×		×	
se8		×		×	×		

a_abondance	Aes	Agb	Aga	Agi	Anc	Ani	Ano	Ant
se1					×			×
se2					×			×
se3	×	×		×		×		×
se4	×	×		×		×		×
se5				×		×	×	
se6				×	×	×		
se7		×	×					×
se8		×					×	×

TAB. 1 – *RCF :* `Taxon`*,* `SiteEchantillon`*,* `a_abondance`*.*

Définition 2 (Concept formel) *Etant donné un contexte objets-attributs $K = (G, M, I)$, un concept associe un ensemble maximal d'objets avec l'ensemble maximal des attributs qu'ils partagent, pour former une paire $C = (Extension(C), Intension(C))$ telle que :*
$Extension(C) = \{g \in G | \forall m \in Intension(C), (g, m) \in I\}$ est l'extension du concept (objets couverts par le concept); $Intension(C) = \{m \in M | \forall g \in Extension(C), (g, m) \in I\}$ est l'intension du concept (attributs partagés).

Les concepts formels sont ordonnés par une relation de spécialisation/généralisation, denotée par $\preceq_C$, fondée sur la relation d'inclusion ensembliste. Etant donnés deux concepts formels $C_1 = (E_1, I_1)$ and $C_2 = (E_2, I_2)$, $C_2 \preceq_C C_1$ si et seulement si $E_2 \subseteq E_1$ (et de façon équivalente $I_1 \subseteq I_2$). C_2 est une spécialisation (i.e., sous-concept) de C_1. C_1 est une généralisation (i.e., super-concept) de C_2. L'intension de C_2 hérite des attributs de l'intension de C_1, tandis que l'extension de C_1 hérite des objets de l'extension de C_2. L'ordre $\preceq_C$ munit l'ensemble des concepts de K d'une structure de treillis, appelée treillis de concepts de K.

La figure 2 montre les treillis de concepts associés aux contextes formels des sites échantillonnés (à gauche) et des taxons (à droite). Le treillis des sites échantillonnés met en évidence le groupe de sites échantillonnés dans des eaux calmes ($C\_SiteEchantillon\_5$) versus le groupe des sites échantillonnés dans des torrents ($C\_SiteEchantillon\_4$). Les sites échantillonnés dans des eaux calmes sont ensuite séparés en trois sous-groupes en fonction de la présence de calcium (`Ca`, pour $C\_SiteEchantillon\_1$), de magnésium (`Mg`, pour $C\_SiteEchantillon\_2$), ou de glyphosate (`C3H8NO5P`, pour $C\_SiteEchantillon\_3$).

L'ARC permet d'introduire des attributs relationnels pour compléter les contextes formels initiaux afin de prendre en compte l'information relationnelle. Une relation $r_j \subseteq G_k \times G_l$ est utilisée pour construire les attributs relationnels de K_k en utilisant les relations entre les objets de G_k et les concepts construits sur les objets de G_l. La figure 3 illustre la notion d'at-

Régler le processus d'exploration dans l'ARC

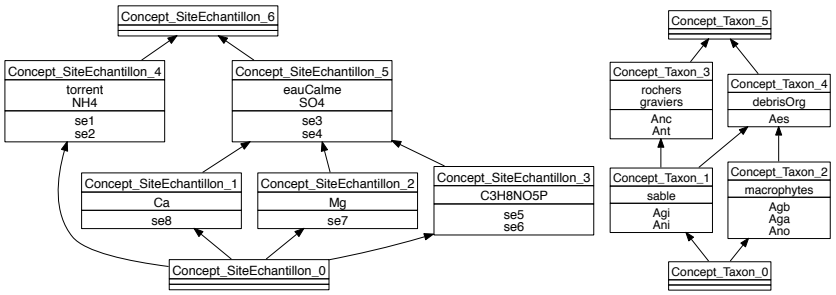

FIG. 2 – *Treillis de concepts pour les contextes formels* `SiteEchantillon` *et* `Taxon`*. Un concept est représenté par une boîte en trois parties. La partie supérieure est son identifiant ; la partie intermédiaire contient l'intension privée des attributs hérités des super-concepts (elle ne contient que les attributs "introduits") ; la partie inférieure contient l'extension privée des objets hérités des sous-concepts (elle ne contient que les objets "introduits").*

tribut relationnel avec quelques exemples. Un attribut relationnel est composé d'un quantifieur d'échelonnage, du nom de la relation, et du concept cible. Par exemple :

- l'attribut relationnel $\exists a\_abondance(Concept\_Taxon\_2)$ est associé aux sites échantillonnés $se3$, $se7$ et $se8$ car ils ont **au moins un** lien *a_abondance* avec un taxon de l'extension du $Concept\_Taxon\_2$.
- l'attribut relationnel $\exists\forall a\_abondance(Concept\_Taxon\_3)$ est associé au site échantillonné $se6$, car il a **au moins un** lien *a_abondance* et ces liens sont **seulement** dirigés vers les taxons de l'extension du $Concept\_Taxon\_3$.
- l'attribut relationnel $\exists\forall_{\geq 60\%} a\_abondance(Concept\_Taxon\_2)$ est associé aux sites échantillonnés $se7$ et $se8$ car ils ont **au moins un** et **au moins 60%** de leurs liens *a_abondance* vers les taxons de l'extension du $Concept\_Taxon\_2$.
- l'attribut relationnel $\exists\supseteq a\_abondance(Concept\_Taxon\_1)$ est associé aux sites échantillonnés $se3$ et $se6$ car ils sont reliés à **au moins un** et à **tous les taxons** de l'extension de $Concept\_Taxon\_1$ à travers un lien *a_abondance*.

En utilisant cette information relationnelle, on peut étendre les contextes formels avec les attributs relationnels et ainsi construire de nouveaux treillis de concepts. Par exemple, à gauche de la figure 4 (resp. à droite) est représenté le treillis de concepts associé au contexte formel des sites échantillonnés étendu avec tous les attributs relationnels possibles composés du quantifieur d'échelonnage $\exists\forall$ (resp. $\exists\forall_{\geq 60\%}$) et des concepts du treillis de concepts des taxons de la figure 2. La figure 4 montre aussi la relation de généralité entre les quantifieurs d'échelonnage. Dans notre exemple, $\exists\forall$ est plus général que $\exists\forall_{\geq 60\%}$ (dénoté par $\exists\forall \preceq_S \exists\forall_{\geq 60\%}$), avec pour conséquence que si un objet possède un attribut relationnel formé avec $\exists\forall$, il possède aussi son équivalent (même relation/même concept) formé avec $\exists\forall_{\geq 60\%}$; il existe ainsi une forme de projection entre les attributs relationnels introduits dans le treillis de gauche vis-à-vis de ceux introduits dans le treillis de droite, et entre les treillis (Braud et al., 2018).

Les questions des hydroécologues pourraient être traitées comme des requêtes à une base de données. Ce que l'AFC apporte est une organisation inhérente des réponses à leurs requêtes par regroupement hiérarchique (Messai et al., 2005; Azmeh et al., 2011). Par exemple, si la

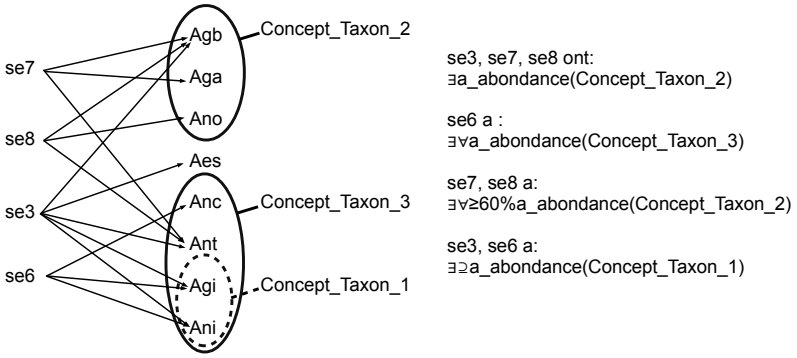

FIG. 3 – *Attributs relationnels construits à partir des concepts* Taxon *et de la relation* a_abondance *entre* SiteEchantillon *et* Taxon

question d'un expert est *trouver les sites échantillonnés en eau calme*, le treillis à gauche de la figure 2 organise les réponses via $C\_SiteEchantillon\_5$ et ses sous-concepts. L'expert voit ainsi quelles réponses correspondent à sa question avec le minimum de caractéristiques additionnelles, alors qu'à mesure qu'il/elle descend dans le treillis, des caractéristiques sont ajoutées aux groupes. Il/elle voit aussi quels sont les sites échantillonnés parmi les réponses "équivalentes" (qui ont exactement les mêmes caractéristiques), ou quelles caractéristiques apparaissent ensemble. Toutes ces informations l'aident à naviguer dans les réponses possibles à sa question, à comprendre des propriétés de ces réponses et à formuler des hypothèses. Grâce à l'ARC, ceci s'étend à l'information relationnelle. Par exemple, le treillis de concepts de la figure 4 (à droite) met en évidence le fait que les sites échantillonnés dans les eaux calmes (C_SE_5) ont une proportion significative de leurs taxons qui apprécient les débris organiques. Il représente plus spécifiquement les réponses à la question générale : "trouver les groupes de sites échantillonnés qui ont plus de 60% de leurs taxons dans un certain groupe de taxons". C_SE_5 et ses sous-concepts montrent l'organisation des réponses à la question *trouver les sites échantillonnés en eau calme* ou de façon alternative *trouver les sites échantillonnés qui ont plus de 60% de leurs taxons ayant pour micro-habitat des débris organiques*. Cette spécificité de l'AFC et de l'ARC porte l'attention de l'expert sur la structuration des données et des relations et l'amène à naviguer dans les réponses et dans le jeu de données.

Dans le cas général, le modèle de données peut être cyclique, impliquant un processus itératif qui converge après un nombre d'étapes dépendant du jeu de données. Par exemple, on peut avoir la relation inverse $est\_abondant$ allant des taxons aux sites échantillonnés. Alors, quand les concepts des sites échantillonnés sont construits, un nouveau treillis de concepts des taxons peut être construit grâce au quantifieur d'échelonnage choisi et aux attributs relationnels en résultant pour $est\_abondant$ et ainsi ouvrir la voie à d'autres questions.

Régler le processus d'exploration dans l'ARC

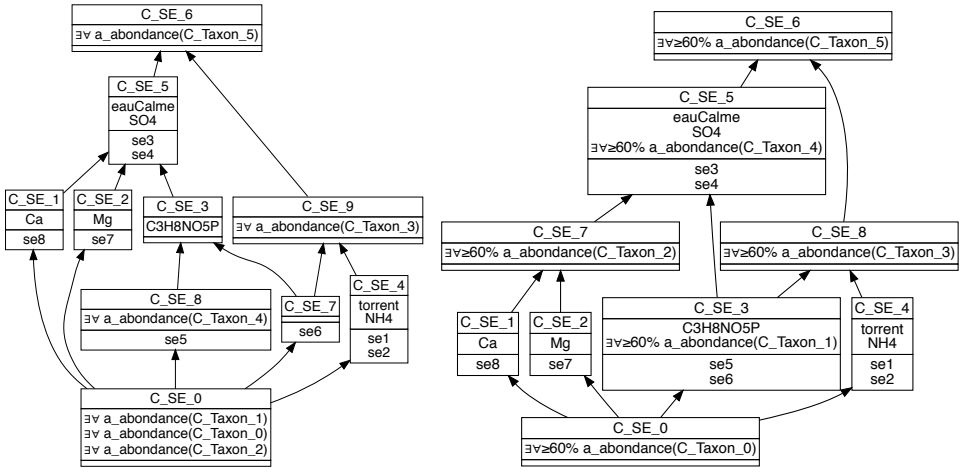

FIG. 4 – *Treillis de concepts SiteEchantillon (SE) avec* $\exists\forall$ *(LHS) et* $\exists\forall_{\geq 60\%}$ *(RHS)*

4 Des guides pour le processus d'exploration de l'ARC

L'outil RCAexplore [2] permet des usages variés de l'ARC : changer à chaque étape le quantifieur d'échelonnage, les contextes formels et les relations considérées, et les concepts calculés. Cette diversité a sa contrepartie dans la difficulté à choisir les bons paramètres pour une question donnée. Pour résoudre cette difficulté, nous avons développé trois sur-couches implantées en python et qui communiquent avec RCAExplore : la première sur-couche concerne l'étape de modélisation, elle permet à l'utilisateur d'exprimer des contraintes sur les relations ; la deuxième et la troisième interviennent à l'étape de choix des quantifieurs permettant la construction des attributs relationnels : l'une pour faciliter l'interprétation des expressions des attributs, l'autre pour anticiper leurs effets, en termes de nombre de concepts et de règles produites. Ces développements ont été faits en proximité avec les hydroécologues du projet Fresqueau, ce qui justifie le caractère appliqué de la présentation que nous en faisons ci-dessous. Notre expérience avec l'ARC nous conduit toutefois à penser que ces outils s'appliqueront de même à d'autres domaines.

Contraindre le choix des quantifieurs d'échelonnage. RCAExplore offre la possibilité de choisir parmi plusieurs quantifieurs pour chaque relation, mais parfois, des relations sont sémantiquement liées et les quantifieurs qui leur sont associés doivent alors être cohérents. Par exemple, dans notre jeu de données, chaque relation générale (p. ex. $a\_abondance$) est représentée par plusieurs relations pour capturer la notion de niveau (p. ex. cinq relations $a\_abondance\_de\_niveau\_i$). Si plusieurs relations $abondance$ sont sélectionnées ensemble, il serait alors cohérent de leur appliquer le même quantifieur. Pour cela, nous regroupons les relations en classes d'équivalence : les relations d'une même classe d'équivalence sont considérées de la même façon au cours du processus et on leur applique le même quantifieur d'éche-

2. http ://dataqual.engees.unistra.fr/logiciels/rcaExplore

lonnage à chaque étape. Néanmoins, le quantifieur d'échelonnage pour une classe peut être différent d'une étape à l'autre. Cette information est analysée quand un quantifieur d'échelonnage est associé à une relation via l'interface utilisateur, afin de propager la contrainte sur les relations de la même classe. L'utilisateur peut accepter ou modifier les propositions du système.

Aide à l'interprétation. Une autre difficulté rencontrée par l'utilisateur est de comprendre l'impact du choix des quantifieurs d'échelonnage sur son analyse. Pour traiter cette difficulté, nous avons développé un interpréteur qui traduit automatiquement les choix réalisés sur l'interface utilisateur en une expression formatée respectant un langage fixé. La partie haute de la figure 5 montre une telle expression, correspondant aux choix faits dans la partie basse. Cette expression est composée à partir d'éléments de la forme :

```
Groupe de <nom d'un Contexte Formel source> qui
<nom d'une Relation> <expression d'un quantifieur >
dans groupe de <nom d'un Contexte Formel cible>
```

Ce type d'expression correspond à un groupe $C = (X, Y) \in \mathcal{L}_{source}$ tel qu'il existe $C' \in \mathcal{L}_{cible}$ avec $qr.C' \in Y$. Le groupe C correspond à `Groupe de <nom d'un Contexte Formel source>` ; r correspond à `<nom d'une Relation>` ; q correspond à `<expression d'un quantifieur>` ; C' correspond à `groupe de <nom d'un Contexte Formel cible>`. Les `<expression d'un quantifieur>` peuvent être :

au moins un <nom d'un Contexte Formel cible>	<nom d'un Contexte Formel cible> seulement
<n> % de <nom d'un Contexte Formel cible>	tous <nom d'un Contexte Formel cible>

La boîte textuelle (en haut de la figure 5) est automatiquement mise à jour quand de nouveaux quantifieurs d'échelonnage sont choisis, clarifiant les choix pour les experts.

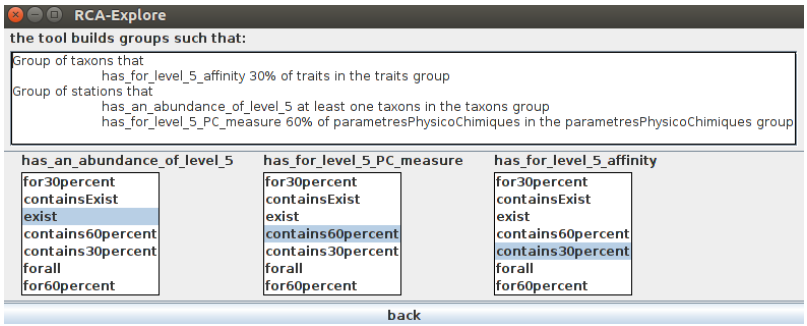

FIG. 5 – *Sélection d'un quantifieur d'échelonnage et interprétation associée (extrait d'écran)*

Tableau de solutions voisines. La troisième sur-couche consiste à calculer des métriques de dimensionnement sur des ensembles de solutions voisines de la requête courante, i.e. qui pourraient être obtenues en variant progressivement les quantifieurs suivant la relation de généralité comme le présentent Braud et al. (2018). Ces métriques peuvent guider l'expert dans la navigation entre les différents réglages des quantifieurs.

Régler le processus d'exploration dans l'ARC

formal context	relation	nb concepts	nb règles	taille maximale support
taxons	∃ affinité 3	460	258	153 sur 1 règle
site-échantillon	∃ mesure 3 ∃ abondance 3	1661	Tot = 419 ma = 19 mm = 6 am = 3 aa = 391	1258 sur 1 règle 2 sur 19 règles 2 sur 6 règles 9 sur 1 règle 1258 sur 1 règle
taxons	∃ affinité 3	460	258	153 sur 1 règle
site-échantillon	∃ mesure 3 $\exists\forall_{\geq 30\%}$ abondance 3	1661	Tot = 434 ma = 32 mm = 3 am = 4 aa = 395	1258 sur 1 règle 2 sur 32 règles 2 sur 3 règles 9 sur 1 règle 1258 sur 1 règle
taxons	∃ affinité 3	460	258	153 sur 1 règle
site-échantillon	∃ mesure 3 $\exists\forall_{\geq 60\%}$ abondance 3	1641	Tot = 472 ma = 11 mm = 3 am = 4 aa = 454	1254 sur 1 règle 2 sur 11 règles 2 sur 3 règles 3 sur 3 règles 1254 sur 1 règle
taxons	∃ affinité 3	460	258	153 sur 1 règle
site-échantillon	∃ mesure 3 ∃∀ abondance 3	1641	Tot = 462 ma = 7 mm = 3 am = 3 aa = 449	930 sur 1 règle 2 sur 7 règles 2 sur 3 règles 3 sur 2 règles 930 sur 1 règle

TAB. 2 – *Exemple de métriques sur les réglages voisins pour ajuster la généralisation des quantifieurs* $universel$ *sur la relation* $a\_abondance$ *(du général au spécifique)*

Dans le tableau 2, nous examinons ce qui se passe quand les quantifieurs (pour l'exemple, $\exists\forall_{\geq n\%}$) sont modifiés sur la relation *a_abondance* de niveau 3. Dans l'étude pratique, pour répondre aux questions sur la connexion entre les paramètres physico-chimiques et les traits de vie des taxons, nous construisons des treillis dans lesquels les experts doivent naviguer, ainsi que des règles. La dimension des résultats se compte en particulier en nombre de concepts des treillis (colonne 3) et certaines règles d'implication non redondantes de prémisse de taille 1 entre attributs relationnels, intéressant les experts (colonne 4). La dernière colonne donne la taille maximale du support des règles (nombre d'objets dans une extension de concept). Dans le treillis des taxons, les règles d'intérêt extraites sont de la forme suivante :
`<affinité quant. d'un groupe traits>` $\Longrightarrow$ `<affinité quant. d'un groupe traits>`.
Dans le treillis des sites échantillonnés, les règles étudiées ont les formes suivantes :
`ma` : `<mesure quant. d'un paramètre PC>` $\Longrightarrow$ `<abondance quant. d'un groupe taxons>`
`aa` : `<abondance quant. d'un groupe taxons>` $\Longrightarrow$ `<abondance quant. d'un groupe taxons>`
`mm` : `<mesure quant. d'un paramètre PC>` $\Longrightarrow$ `<mesure quant. d'un paramètre PC>`
`am` : `<abondance quant. d'un groupe taxons>` $\Longrightarrow$ `<mesure quant. d'un paramètre PC>`

Une règle `ma` révèle un lien entre l'état physico-chimique du cours d'eau et le niveau de présence de certains groupes de taxons. Une règle `aa` révèle la présence simultanée de taxons. Ces deux types de règles sont cohérents avec les questions des hydroécologues. Une règle `mm` donne des résultats déjà connus, comme p. ex. des relations entre les différentes formes de l'azote, qui dépendent des processus chimiques. Les règles `am` sont a priori moins pertinentes car les taxons (ici macroinvertébrés) n'ont pas d'effet attendu sur les paramètres physico-chimiques. L'outil offre la possibilité de ne pas calculer les types de règles, comme

ce dernier, s'ils n'intéressent pas l'expert. En exploitant le tableau 2, l'expert peut d'abord remarquer que le jeu de données contient des échantillons très diversifiés en terme de physico-chimie et de population de taxons, ce qui explique la faible valeur du support maximal des règles en général. Si l'expert est intéressé par les classifications, il remarque qu'il n'y pas de grande différence quand le quantifieur d'échelonnage est spécialisé (de 1661 concepts à 1641). Il note que les treillis obtenus pour $\exists\forall_{\geq 60\%}$ et $\exists\forall$ ont le même nombre de concepts ; il est donc inutile de raffiner ce quantifieur jusqu'à 100% si le but est de réduire la taille des résultats pour en faciliter l'analyse. Si l'expert s'intéresse aux règles `mesure/abondance (ma)`, il peut en examiner le nombre dans la colonne 4, successivement 19, 32, 11 et 7. Il peut alors décider d'utiliser le quantifieur $\exists\forall_{\geq 60\%}$ qui donne un nombre raisonnable de règles reposant sur une connexion relativement élevée entre les mesures physico-chimiques et les taxons.

5 Conclusion

Dans cet article, nous avons présenté plusieurs sur-couches de l'outil en accès libre RCAExplore de manière à guider des experts métier, qui n'auraient pas une connaissance approfondie de l'ARC, dans l'exploration de leurs données. L'ARC permet d'extraire des classifications, des règles et des motifs avec une grande variété de filtres procurés par les quantifieurs, mais les modifications apportées aux résultats par ces quantifieurs peuvent ne pas être très intuitives et le nombre de concepts générés peut se révéler important. L'application des contraintes permet de prendre en compte des groupes cohérents de relations et de leur appliquer des quantifieurs de manière homogène. L'interprète de requêtes inclus dans l'outil permet de les clarifier et aide à la fois à les formuler et à comprendre les résultats obtenus. Les scripts calculent des métriques et extraient des règles d'implication respectant certaines formes. Les métriques permettent aux experts de disposer d'un aperçu sur les treillis (nombre de concepts, nombre et support des règles générées), et ainsi de réorienter l'analyse, en étendant ou en restreignant la recherche. Les règles d'implication informent les experts sur les relations entre objets du domaine.

Pour le futur, nous projetons de développer une version de RCAExplore qui soit un outil d'exploration de données totalement intégré permettant de passer des données brutes à des résultats construits exploitables par les experts métier. Pour cela, une interface devra rendre transparent l'enchaînement de l'appel des différentes parties de l'outil, mettre à disposition tous les résultats, certains étant actuellement stockés dans des fichiers et présenter un menu pour faciliter le processus itératif. Les langages de contraintes et d'interprétation mériteront d'être affinés pour se rapprocher du langage naturel, de manière à procurer une interface plus simple d'utilisation. Nous devrons aussi analyser la réaction des utilisateurs à cette interface et voir en quoi elle facilite effectivement l'usage de l'ARC.

Enfin il sera pertinent de travailler sur la complémentarité entre l'ARC et d'autres approches. Nous avons déjà réalisé une comparaison avec la recherche de motifs temporels et l'apprentissage inductif ; d'autres travaux existent sur les liens entre l'ARC et le bi-clustering, ou, plus généralement, l'apprentissage (Kaytoue et al., 2015). Ces comparaisons pourront être étendues aux bases de données relationnelles, aux bases de données inductives, et également à OWL/RDF et SPARQL.

Remerciements. Ce travail a été partiellement financé par l'AFB (Agence française de biodiversité). Merci à Corinne Grac (UMR 7362 LIVE - ENGEES) pour son aide.

Références

Azmeh, Z., M. Huchard, A. Napoli, M. R. Hacene, et P. Valtchev (2011). Querying relational concept lattices. In *Proceedings of CLA 2011*, pp. 377–392.

Braud, A., X. Dolques, M. Huchard, et F. Le Ber (2018). Generalization effect of quantifiers in a classification based on relational concept analysis. *Know.-Based Syst. 160*(15), 119–135.

Dolques, X., F. Le Ber, M. Huchard, et C. Nebut (2015). Relational Concept Analysis for Relational Data Exploration. *Adv. in Knowledge Discovery and Management 5*, 55–77.

Džeroski, S. (2003). Multi-relational data mining : an introduction. *ACM SIGKDD Explorations Newsletter 5*(1), 1–16.

Ferré, S. (2015). A Proposal for Extending Formal Concept Analysis to Knowledge Graphs. In *Formal Concept Analysis, ICFCA 2015*, Volume LNCS 9113, Nerja, Spain, pp. 271–286.

Ferré, S., O. Ridoux, et B. Sigonneau (2005). Arbitrary Relations in Formal Concept Analysis and Logical Information Systems. In *ICCS'05*, LNAI 3596, pp. 166–180. Springer.

Ganter, B. et R. Wille (1999). *Formal Concept Analysis : Math. Foundations*. Springer Verlag.

Hacene, M. R., M. Huchard, A. Napoli, et P. Valtchev (2013). Relational concept analysis : mining concept lattices from multi-relational data. *Ann. Math. Artif. Intell. 67*(1), 81–108.

Kaytoue, M., V. Codocedo, A. Buzmakov, J. Baixeries, S. O. Kuznetsov, et A. Napoli (2015). Pattern structures and concept lattices for data mining and knowledge processing. In *Machine Learning and Knowledge Discovery in Databases*, pp. 227–231. Springer.

Kötters, J. (2013). Concept Lattices of a Relational Structure. In *ICCS 2013, Mumbai, India*, LNCS 7735, pp. 301–310.

Liquière, M. et J. Sallantin (1998). Structural Machine Learning with Galois Lattice and Graphs. In *ICML, Madison, Wisconsin*, pp. 305–313.

Messai, N., M. Devignes, A. Napoli, et M. Smaïl-Tabbone (2005). Querying a bioinformatic data sources registry with concept lattices. In *Conceptual Structures : Common Semantics for Sharing Knowledge, ICCS 2005, Proceedings*, pp. 323–336.

Palagi, É., F. L. Gandon, A. Giboin, et R. Troncy (2017). A survey of definitions and models of exploratory search. In *ACM Workshop ESIDA@IUI*, pp. 3–8.

Wildemuth, B. M. et L. Freund (2012). Assigning search tasks designed to elicit exploratory search behaviors. In *Human-Computer Inf. Retriev. Symp. (HCIR)*.

Summary

This paper focuses on the exploration of multi-relational datasets, and the various ways they can be analyzed using relational concept analysis (RCA), a variant of Formal Concept Analysis. RCA uses several scaling operators that make the process highly tunable, allowing a high flexibility in the exploration and in the results. Besides, the multiplicity of choices that can be made when performing an analysis task is potentially overwhelming the expert. We therefore propose three overlays for helping users controlling, and foreseeing the results of their choices. This proposition is exemplified on a dataset about the quality state of watercourses.