L'analyse relationnelle de concepts pour la fouille de données temporelles – Application à l'étude de données hydroécologiques

Cristina Nica*, Agnès Braud*, Xavier Dolques*
Marianne Huchard**, Florence Le Ber*

*ICube, Université de Strasbourg, CNRS, ENGEES
prenom.nom@engees.unistra.fr, agnes.braud@unistra.fr
http://icube-bfo.unistra.fr
**LIRMM, Université de Montpellier, CNRS
huchard@lirmm.fr
https://www.lirmm.fr

Résumé. Cet article présente une méthode d'exploration de données temporelles, fondée sur l'analyse relationnelle de concepts (ARC) et appliquée à des données séquentielles construites à partir d'échantillons physico-chimiques et biologiques prélevés dans des cours d'eau. Notre but est de mettre au jour des sous-séquences pertinentes et hiérarchisées, associant les deux types de paramètres. Pour faciliter la lecture, ces sous-séquences sont représentées sous la forme de motifs partiellement ordonnés (po-motifs). Le processus de fouille de données se décompose en plusieurs étapes : construction d'un modèle temporel *ad hoc* et mise en œuvre de l'ARC ; extraction des sous-séquences synthétisées sous la forme de po-motifs ; sélection des po-motifs intéressants grâce à une mesure exploitant la distribution des extensions de concepts. Le processus a été testé sur un jeu de données réelles et évalué quantitativement et qualitativement.

1 Introduction

Les données temporelles ou spatio-temporelles, de plus en plus nombreuses et largement utilisées, ont conduit au développement d'un ensemble de méthodes, permettant l'extraction de motifs ordonnés fréquents, et relevant de la « fouille de données séquentielles » (Agrawal et Srikant, 1995). Considérons par exemple une base de données de maladies d'où sont extraits deux motifs séquentiels fréquents $\langle$(`Migraine`)(`Fièvre`)$\rangle$ et $\langle$(`Gorge-enflammée`)(`Fièvre`)$\rangle$. Ces deux motifs mettent en évidence les symptômes qui précédent la fièvre, mais leur coexistence n'est pas prise en compte dans les approches classiques. Pour surmonter cette limitation, Casas-Garriga (2005) a proposé d'utiliser des motifs partiellement ordonnés (po-motifs). Les po-motifs ont pour intérêt d'être compacts, de contenir la même information que les ensembles de motifs qu'ils représentent, et d'être faciles à interpréter grâce à leur representation sous forme de graphes orientés acycliques. Les motifs partiellement ordonnés ont déjà été étudiés selon différentes méthodes, telles que Orderspan (Fabrègue et al., 2015) et Frecpo

(Pei et al., 2006). Ces approches sont efficaces, mais ont pour défaut d'extraire un ensemble de po-motifs sans ordre, ce qui rend l'étape d'analyse très difficile sans post-traitement.

Cet article propose une nouvelle méthode de fouille de séquences de données qualitatives pouvant contenir des répétitions. Nous proposons d'utiliser l'analyse relationnelle de concepts (ARC, (Rouane-Hacene et al., 2013)), qui classe des ensembles d'objets décrits par des attributs et des relations, permettant ainsi de découvrir des répétitions et des règles d'implication dans des ensembles de données relationnelles. L'ARC a été appliquée dans des champs variés, par exemple pour l'analyse de logiciels et la ré-ingénierie (Arévalo et al., 2006). L'ARC produit pour résultat une famille de treillis de concepts, où chaque treillis peut avoir un grand nombre de concepts. En conséquence, pour faciliter la phase d'analyse de ces résultats, il est nécessaire de disposer de procédures permettant la sélection de concepts pertinents. Notre objectif est 1) d'étendre l'ARC pour extraire des po-motifs, en s'appuyant sur sa capacité à classer des données relationnelles et à produire des résultats hiérarchisés, 2) de montrer que l'ARC permet d'explorer des données temporelles qualitatives, 3) de faciliter l'analyse des résultats grâce à l'ordre de généralisation-spécification intrinsèque à l'ARC, 4) de proposer une méthode de sélection des po-motifs en utilisant les différentes granularités des objets temporels.

Dans ce but, nous proposons un *modèle temporel des données* mettant en exergue un *treillis cible* (contenant les objets à analyser). Chaque concept du treillis cible correspond à un ensemble de sous-séquences temporelles qui sont synthétisées sous la forme d'un po-motif. Sur cette base, nous procédons de la façon suivante (voir figure 1). D'abord, nous appliquons l'ARC sur une *famille de contextes relationnels* contenant les données. Ensuite nous extrayons les sous-séquences à partir des concepts du treillis cible et nous construisons les po-motifs. Finalement, nous choisissons les motifs pertinents en calculant des mesures d'interêt, fondées sur la *distribution* des extensions de concepts. De plus, grâce à la structure hiérarchique fournie par l'ARC, les motifs sont classés du plus général au plus spécifique, permettant à l'analyste de naviguer facilement dans l'espace des po-motifs sans calcul supplémentaire.

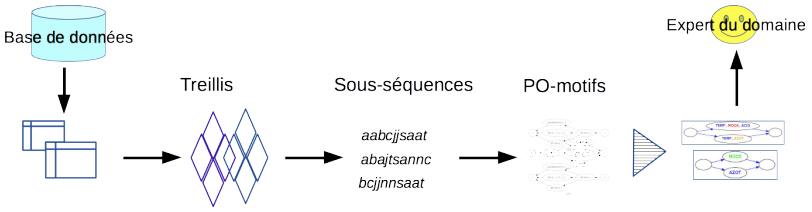


FIG. 1: Schéma du processus d'analyse.

L'approche et les résultats présentés ici reposent sur des jeux de données collectés dans le cadre du projet ANR 11 MONU 14 Fresqueau [1], qui s'intéressait à l'évaluation de la qualité des cours d'eau. Les données représentent des mesures biologiques et physico-chimiques effectuées en des points fixes sur des cours d'eau (appelés stations) et répétées dans le temps. Les prétraitements appliqués sur ces données sont décrits dans (Nica et al., 2015).

L'article est organisé de la façon suivante. La section 2 est un état de l'art. Les sections 3, 4 et 5 introduisent l'ARC et les différentes étapes de notre approche, illustrées par un petit extrait

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

des jeux de données de Fresqueau. La section 6 présente et discute les résultats expérimentaux obtenus sur les jeux de données complets. La dernière section est une conclusion.

2 Etat de l'art

À notre connaissance, l'ARC n'a jamais été utilisée pour explorer des données temporelles. Il existe néanmoins plusieurs approches voisines, fondées sur l'analyse de concepts formels (ACF, (Ganter et Wille, 1999)). Wolff (2001) a introduit l'analyse de concepts temporels, où les objets sont caractérisés par une date et un état (un ensemble d'attributs). Les données sont représentées dans un unique contexte et le treillis de concepts résultant est analysé grâce à l'élément date présent dans les concepts, révélant les relations temporelles entre concepts. Cette approche a été utilisée pour analyser des données séquentielles concernant des activités suspectes (Poelmans et al., 2010). Dans notre approche fondée sur l'ARC, la relation temporelle entre les dates est considérée comme une relation objet-objet et elle lie les concepts de différents treillis. Ferré (2007) utilise des arbres de suffixes pour rechercher des sous-chaînes maximales. Buzmakov et al. (2015) transforment les données séquentielles en une structure de motifs munie d'un ordre partiel, permettant de construire un treillis de concepts. Les auteurs combinent la mesure de stabilité des concepts et des opérations de projection pour sélectionner les motifs intéressants.

Il existe plus largement de nombreuses méthodes pour explorer les données séquentielles. Casas-Garriga (2005) extrait des sous-séquences fermées qui sont ensuite transformées en po-motifs puis organisées dans un treillis similaire à un treillis de concepts. Dans notre approche, les po-motifs extraits sont déjà hiérarchisés. L'algorithme Frecpo, développé par Pei et al. (2006) pour explorer des bases de chaînes de caractères, permet d'extraire des po-motifs à partir de séquences simples sans répétition, alors que notre approche permet de traiter des séquences contenant des répétitions (comme pour les données utilisées ci-dessous). Une telle approche a également été développée dans le projet Fresqueau, il s'agissait plus précisément de rechercher des po-motifs fermés, triés ensuite à l'aide de différentes mesures (Fabrègue et al., 2014). Dans l'approche présentée ici, les po-motifs sont filtrés par des mesures exploitant les caractéristiques spatiales des données.

3 Analyse relationnelle de données temporelles

Nous nous intéressons à des jeux de données séquentielles issues de mesures biologiques (Bio) et physico-chimiques (PhC) effectuées dans des rivières. Une mesure est identifiée par un *objet temporel*, soit une paire *(objet, date)* où *objet* est un objet pérenne, ici une station de rivière, et $date$ est un point temporel où l'état de l'objet est observé, ici la date à laquelle la mesure a été effectuée. Un ensemble de mesures effectuées sur la même station et ordonnées chronologiquement constitue une séquence de données.

Le jeu de données séquentielles est structuré selon le schéma décrit en figure 2. Les quatre rectangles représentent les quatre ensembles d'objets manipulés : mesures Bio, mesures PhC, paramètres Bio et paramètres PhC. Les mesures sont liées par une relation binaire temporelle `précédé par`, qui associe une mesure à une autre si la première est précédée dans le temps par la seconde sur la même station. Les mesures Bio (respectivement PhC) sont liées aux

paramètres Bio (respectivement PhC) par plusieurs relations binaires qualitatives qui associent une mesure à un paramètre si la qualité de cette mesure concernant le paramètre est la qualité représentée par la relation. Nous cherchons à évaluer l'influence des paramètres PhC (p. ex. azote (AZOT), phosphore (PHOS), particules en suspension (PAES)) sur les paramètres Bio (p. ex. l'indice biologique diatomées (IBD) et l'indice biologique global normalisé (IBGN)).

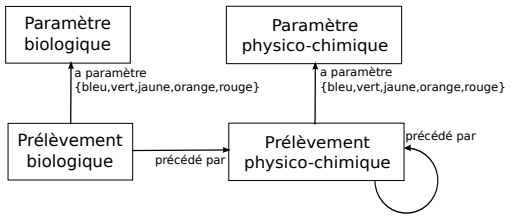


FIG. 2: Schéma relationnel illustrant la modélisation du jeu de données séquentielles ; les couleurs `bleu`, `vert`, `jaune`, `orange` et `rouge` représentent respectivement les classes de qualité très bonne, bonne, moyenne, mauvaise et très mauvaise des paramètres.

L'ARC étend le cadre de l'ACF aux données relationnelles. L'ACF considère un contexte formel qui est un ensemble d'objets décrits par des attributs, et construit un treillis de concepts qui permet d'analyser les objets. Brièvement, un contexte formel K est un triplet (G, M, I), où G est un ensemble d'objets, M un ensemble d'attributs, et I la relation d'incidence, $I \subseteq G \times M$. Un concept formel issu de K est un couple $C = (X, Y)$, où $X = \{g \in G | \forall m \in Y, (g, m) \in I\}$ et $Y = \{m \in M | \forall g \in X, (g, m) \in I\}$; X et Y sont respectivement l'extension et l'intension du concept. Soit $\mathcal{C}_K$ l'ensemble de tous les concepts formels issus de K. Soient $C_1 = (X_1, Y_1)$ et $C_2 = (X_2, Y_2)$ deux concepts de $\mathcal{C}_K$, l'ordre de généralisation sur les concepts $\preceq_{\mathcal{C}_K}$ est ici défini par $C_1 \preceq_{\mathcal{C}_K} C_2$ si et seulement si $X_1 \subseteq X_2$ (qui est équivalent à $Y_2 \subseteq Y_1$). On note $\mathcal{L}_K$=($\mathcal{C}_K$, $\preceq_{\mathcal{C}_K}$) le treillis de concepts issu de K. Dans ce treillis, $\top(\mathcal{L}_K)$ est le concept dont l'extension contient tous les objets de G et $\bot(\mathcal{L}_K)$ le concept de $\mathcal{L}_K$ dont l'intension contient tous les attributs de M.

contextes objet-attribut

Kphc	AZOT	PHOS	PAES
AZOT	×		
PHOS		×	
PAES			×

Kphcs
S1_07/07
S1_06/08
S2_04/08
S2_03/08
S2_08/04
S2_07/04
S1_07/08
S3_12/07
S2_06/06

relations objet-objet

RPhC-pp-PhC	S1_07/07	S1_06/08	S2_04/08	S2_03/08	S2_08/04	S2_07/04	S1_07/08	S3_12/07	S2_06/06
S1_07/07									
S1_06/08									
S2_04/08				×					
S2_03/08									
S2_08/04						×			
S2_07/04									
S1_07/08		×							
S3_12/07									
S2_06/06									

RbPhC	AZOT	PHOS	PAES
S1_07/07	×		
S1_06/08	×		
S2_04/08		×	×
S2_03/08			
S2_08/04	×		×
S2_07/04	×		×
S1_07/08	×		×
S3_12/07	×		
S2_06/06			

RvPhC	AZOT	PHOS	PAES
S1_07/07			×
S1_06/08			×
S2_04/08	×		
S2_03/08			
S2_08/04		×	
S2_07/04			
S1_07/08		×	
S3_12/07			×
S2_06/06			

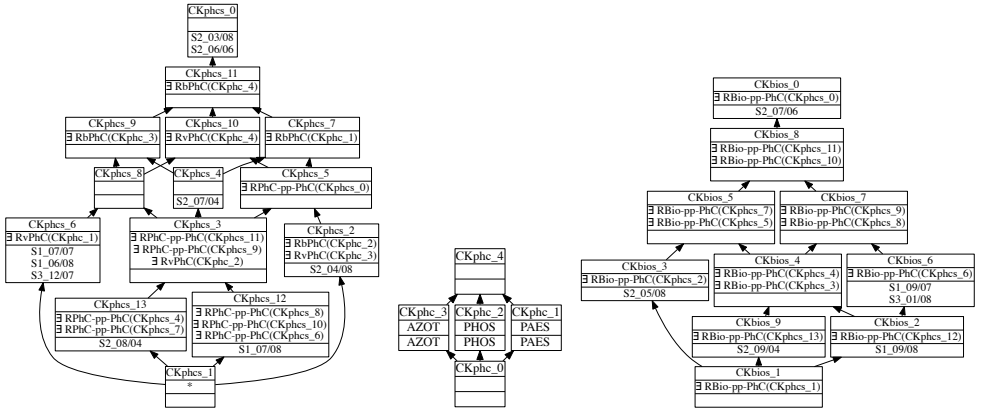
TAB. 1: FRC composée des contextes objet-attribut `Kphc` (paramètres PhC) et `Kphcs` (mesures PhC), avec une relation objet-objet temporelle (précédé par) `RPhC-pp-PhC` et des relations objet-objet qualitatives (a paramètre bleu) `RbPhC` et (a paramètre vert) `RvPhC`.

L'ARC est une mise en œuvre itérative de l'ACF sur une famille relationnelle de concepts (FRC). Une FRC est constituée d'un ensemble $\mathcal{K}$ de contextes objet-attribut et d'un ensemble $\mathcal{R}$ de contextes objet-objet. Un exemple de FRC sur un petit sous-ensemble de nos données est

présenté au tableau 1. Elle est définie pour une seule valeur de paramètre Bio, ici le paramètre *IBGN*$_{bleu}$. Pour simplifier, nous présentons seulement la partie physico-chimique des données. Le modèle complet utilisé pour les expérimentations peut être déduit de la figure 2 et le lecteur peut se référer à Nica et al. (2015) pour obtenir des informations détaillées sur le jeu de données. $\mathcal{K}$ contient n contextes formels objet-attribut $K_i = (G_i, M_i, I_i), i \in \{1, ..., n\}$. Dans le tableau 1 les deux tableaux à gauche représentent les contextes objet-attribut : chaque ligne désigne un objet, chaque colonne un attribut et chaque croix indique que l'objet de la ligne et l'attribut de la colonne forment un couple de la relation d'incidence. $\mathcal{R}$ contient m contextes relationnels objet-objet $R_j = (G_k, G_l, r_j), j \in \{1, ..., m\}$, où G_k que nous appelons le domaine de la relation et G_l que nous appelons le co-domaine de la relation sont respectivement les ensembles d'objets de K_k et K_l, et $r_j \subseteq G_k \times G_l, k, l \in \{1, ..., n\}$. Dans le tableau 1, les trois tableaux à droite représentent les contextes objet-objet : chaque ligne désigne un objet du domaine de la relation, chaque colonne est un objet du co-domaine de la relation et chaque croix indique un lien entre deux objets. `RPhC-pp-PhC` décrit la relation temporelle entre les mesures PhC avec l'ensemble des objets de `Kphcs` comme domaine et co-domaine. `RbPhC` et `RvPhC` décrivent une relation de qualité entre les mesures et les paramètres PhC, elles représentent respectivement les qualités bleue et verte.

On note ici que le contexte `Kphcs` n'a pas de colonne, ce qui signifie que son ensemble d'attributs est vide. Les mesures sont donc décrites uniquement par les relations objet-objet. De nouveaux attributs, appelés *attributs relationnels*, étendent les contextes formels à partir des relations objet-objet et des concepts déjà créés. Un attribut relationnel prend la forme syntaxique $qr_j(C)$, où q est un quantifieur, r_j est une relation et C un concept dont l'extension contient des objets du co-domaine de r_j. Ici nous utilisons le quantifieur *existentiel* qui, pour une relation $R_j = (G_k, G_l, r_j)$, crée une relation $\exists r_j$ entre un objet $o \in G_k$ et un concept $C = (X, Y)$ du treillis $\mathcal{L}_{K_l}$ si $r_j(o) \cap X \neq \emptyset$. Par exemple, l'attribut relationnel $\exists$`RvPhC`(`CKphc_1`) est un attribut commun à toutes les mesures pour lesquelles le paramètre représenté par `CKphc_1` (`PAES`, figure 3 (b)) est mesuré avec la qualité verte. Le processus ARC consiste en l'application de l'ACF d'abord sur chaque contexte objet-attribut d'une FRC, et ensuite, sur chaque contexte objet-attribut étendu par les attributs relationnels créés en utilisant les concepts de l'étape précédente. Le processus s'arrête quand les familles de treillis obtenues pour deux étapes consécutives sont isomorphes et que les contextes n'ont pas changé.

La figure 3 présente la famille de treillis de concepts obtenue par l'application de l'ARC sur la FRC du tableau 1. Chaque concept est représenté par une boîte faite de trois rectangles : le rectangle haut contient le nom du concept ; le rectangle intermédiaire contient l'intension simplifiée et le rectangle bas l'extension simplifiée du concept. La représentation de chaque treillis est simplifiée car tout attribut (resp. objet) est hérité de haut en bas (resp. de bas en haut). Ainsi, un attribut (resp. objet) n'est affiché que dans le concept le plus haut (resp. le plus bas) dans lequel il apparaît. Les flèches représentent l'ordre de généralisation. Le treillis à gauche est le treillis des mesures PhC, $\mathcal{L}_{Kphcs}$. Ses concepts peuvent être interprétés en suivant les références à d'autres concepts du même treillis ou du treillis des paramètres PhC, $\mathcal{L}_{Kphc}$ (au milieu). Par exemple, `CKphcs_6` contient les mesures de qualité verte pour `PAES` ($\exists$`RvPhC`(`CKphc_1`)) et de qualité bleue pour `AZOT` ($\exists$`RbPhC`(`CKphc_3`) hérité de `CKphcs_9`). $\exists$`RvPhC`(`CKphc_4`) et $\exists$`RbPhC`(`CKphc_4`) sont impliqués par les attributs relationnels précédents par héritage dans le treillis $\mathcal{L}_{Kphc}$. Le treillis à droite est le treillis des prélèvements Bio, $\mathcal{L}_{Kbios}$, qui est interprété en liaison avec le treillis $\mathcal{L}_{Kphcs}$.



(a) $\mathcal{L}_{Kphcs}$ (b) $\mathcal{L}_{Kphc}$ (c) $\mathcal{L}_{Kbios}$ (cible)

FIG. 3: Famille de treillis de concepts obtenue en appliquant l'ARC sur le jeu de données : (a) et (b) sont obtenus à partir de la FRC présentée au tableau 1 ; le symbole $*$ représente l'ensemble des attributs de `Kphcs`.

4 Extraction de po-motifs à partir des treillis relationnels

L'extraction des po-motifs est réalisée en deux étapes : 1) construction de sous-séquences à partir des concepts et 2) transformation d'ensembles de sous-séquences en po-motifs.

Soit $\mathcal{I} = \{I_1, I_2, ..., I_m\}$ un ensemble d'*items*. On appelle *itemset* un ensemble d'*items* IS non vide, non ordonné, $IS = (I_{j_1}...I_{j_k})$ où $I_{j_i} \in \mathcal{I}$. Soit $\mathcal{IS}$ l'ensemble de tous les *itemsets* issus de $\mathcal{I}$. Une séquence S est une liste ordonnée non vide d'*itemsets*, $S = \langle IS_1 IS_2 ... IS_p \rangle$ où $IS_i \in \mathcal{IS}$ pour tout $i \in [1, p]$. La séquence S est une sous-séquence d'une autre séquence $S' = \langle IS'_1 IS'_2 ... IS'_q \rangle$, noté $S \preceq_s S'$, si $p \leq q$ et s'il existe des entiers $j_1 < j_2 < ... < j_k < ... < j_p$ tels que $IS_1 \subseteq IS'_{j1}, IS_2 \subseteq IS'_{j2}, ..., IS_p \subseteq IS'_{jp}$.

Dans ce travail un *item* d'une séquence est un paramètre PhC doté d'une valeur qualitative, p.ex. $AZOT_{bleu}$. On construit une sous-séquence en naviguant au long d'une chaîne de concepts reliés par des relations temporelles. L'interprétation de l'intension d'un concept de la chaîne permet de définir : 1) un *itemset* dont les éléments sont les *attributs relationnels qualitatifs* qui dérivent des relations qualitatives et 2) la position de l'*itemset* (dans la sous-séquence) déterminée par les *attributs relationnels temporels* qui dérivent des relations temporelles. On note qu'il y a au moins une sous-séquence pour chaque attribut relationnel temporel dans l'intension d'un concept du treillis cible, c'est-à-dire qu'un ensemble de sous-séquences peut *a priori* être extrait pour chaque concept de ce treillis, sauf si ce concept n'admet pas d'attribut relationnel temporel, auquel cas l'ensemble de sous-séquences est vide.

Une approche naïve d'interprétation de l'intension d'un concept prendrait en compte tous ses attributs relationnels qualitatifs et temporels. Nous proposons une approche plus efficace pour extraire des sous-séquences d'une famille de treillis, en nous appuyant sur la relation d'ordre $\preceq_{\mathcal{C}_K}$ entre les concepts utilisés pour construire les attributs relationnels qualitatifs et temporels. Soient C_1 et C_2 deux concepts du treillis $\mathcal{L}_K$ tels que $C_1 \preceq_{\mathcal{C}_K} C_2$. Si l'intension d'un concept C contient deux (ou plus) attributs relationnels $\exists r(C_1)$ et $\exists r(C_2)$ (issus de la

même relation r), alors $\exists r(C_1) \rightarrow \exists r(C_2)$. Ainsi, $\exists r(C_2)$ est redondant pour l'interprétation de C. En exploitant les attributs relationnels portant sur les concepts les plus spécifiques, on construit alors une interprétation minimalement redondante de l'intension d'un concept.

Par ailleurs, trois règles principales guident l'extraction des ensembles de sous-séquences : 1) une intension de concept vide est redondante, 2) une intension d'un concept qui ne contient que des attributs relationnels temporels n'apporte pas d'information et 3) pour un concept C_3 et tous les couples (C_1, C_2) tels que $\{\exists r(C_1), \exists r(C_2)\} \subseteq intent(C_3)$, où r est une relation temporelle RBio-pp-PhC ou RPhC-pp-PhC, si $\exists$RPhC-pp-PhC$(C_1) \in intent(C_2)$ alors $\exists r(C_1)$ est redondant pour l'interprétation de C_3.

Considérons la sous-séquence extraite pour CKbios_4 (figure 3 (c)) qui est construite à partir de l'attribut relationnel temporel $\exists$RBio-pp-PhC(CKphcs_3). En analysant l'intension du concept CKphcs_3 (figure 3 (a)), on voit que le concept le plus spécifique suivant et final (son intension n'a pas d'attribut relationnel temporel) est CKphcs_9. Ainsi, la chaîne de concepts extraite est $\langle$(CKphcs_3)(CKphcs_9)$\rangle$. Dans ce cas, pour le concept CKbios_4, on n'extrait qu'une sous-séquence notée S_{CKbios_4}. Pour définir les *itemsets* de S_{CKbios_4}, les attributs relationnels qualitatifs des intensions de CKphcs_3 et CKphcs_9 doivent être interprétés. Dans ce but, nous définissons deux types d'attributs relationnels, selon le caractère spécifique ou général de leur interprétation.

Definition 1 (**Attribut relationnel vague/défini**). *L'attribut relationnel $\exists r(C)$, où C est un concept du treillis $\mathcal{L}_K$, est dit* vague *ou* général *si $C \equiv \top(\mathcal{L}_K)$. Il est dit* défini *ou* spécifique *si $C \prec_{\mathcal{C}_K} \top(\mathcal{L}_K)$.*

Ces deux types d'attributs relationnels qualitatifs mettent en évidence des valeurs PhC générales (vagues) ou spécifiques (définies) liées à l'extension des concepts impliqués dans les attributs. Ici nous nous focalisons sur des ensembles de sous-séquences contenant des valeurs PhC spécifiques, extraites des concepts dont les intensions contiennent des attributs relationnels qualitatifs définis. Dans la figure 3 (a), l'intension de CKphcs_9 contient les deux types d'attributs relationnels qualitatifs, le vague $\exists$RbPhC(CKphc_4) et le défini $\exists$RbPhC(CKphc_3). On vérifie CKphc_3 $\prec_{\mathcal{C}_K}$ CKphc_4, ainsi CKphcs_9 est interprété *via* $\exists$RbPhC(CKphc_3) qui pointe de manière univoque vers l'occurrence d'un paramètre PhC bleu (AZOT, dans l'extension de CKphc_3, figure 3 (b)) noté $AZOT_{bleu}$. Selon le même principe, l'interprétation de CKphcs_3, notée $PAES_{bleu}$ $AZOT_{bleu}$ $PHOS_{vert}$, est obtenue *via* les trois attributs relationnels définis présents dans l'intension de CKphcs_3, $\exists$RbPhC(CKphc_1), $\exists$RbPhC(CKphc_3) et $\exists$RvPhC(CKphc_2), révélant trois paramètres PhC mesurés à la même date. La sous-séquence extraite est alors $S_{CKbios\_4} = \langle(\text{PAES}_{\text{bleu}}\ \text{AZOT}_{\text{bleu}}\ \text{PHOS}_{\text{vert}})(\text{AZOT}_{\text{bleu}})\rangle$. La figure 4 montre la navigation nécessaire à l'extraction de cette séquence, S_{CKbios_4}, par l'approche naïve et par celle que nous avons proposée, la seconde permettant de limiter l'analyse à 6 (en gras dans la figure) parmi 14 concepts.

Les ensembles de sous-séquences extraits des treillis de l'ARC sont ensuite convertis en po-motifs. Pour cela, nous utilisons les procédures d'élagage et de fusion décrites par Fabrègue et al. (2015). Formellement, un *po-motif* est un graphe orienté acyclique $G = (\mathcal{V}, \mathcal{E})$. $\mathcal{V}$ est l'ensemble des sommets et $\mathcal{E}$ l'ensemble des arcs tel que $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$. Cette structure permet de fixer un ordre partiel sur les sommets u et v, défini par $u < v$ s'il existe un chemin direct de u vers v. Si ce chemin n'existe pas, u et v sont dits incomparables. Un po-motif G est associé à l'ensemble de sous-séquences S qui contiennent tous les chemins de G. Le po-motif obtenu à partir de la séquence S_{CKbios_4} est montré en figure 5.

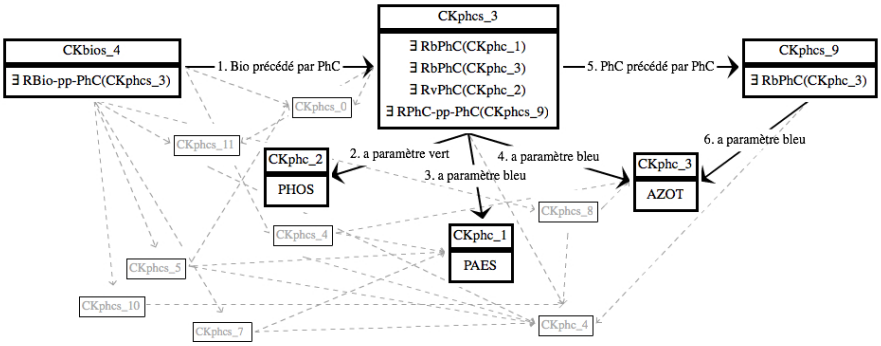


FIG. 4: Ensemble des intensions/extensions de concepts utilisées pour interpréter `CKbios_4`.

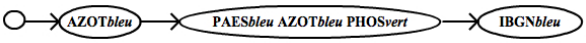


FIG. 5: Le po-motif extrait de `CKbios_4`.

5 Filtrage des po-motifs

Pour faciliter le travail de l'analyste face au grand nombre de po-motifs extraits, il faut définir des méthodes de sélection des po-motifs intéressants. Le terme « intéressant » peut avoir différentes significations selon le domaine et les objectifs de l'analyste. Les mesures généralement utilisées pour sélectionner les concepts, telles que la stabilité (Kuznetsov, 2007), la probabilité et la séparation (Klimushkin et al., 2010), ne sont pas efficaces ici, d'après les experts du domaine, pour déterminer si un po-motif est intéressant. Pour eux, un po-motif intéressant doit être fréquent et associé à un grand ensemble de stations échantillonnées dans les mêmes proportions. En conséquence, nous proposons ci-dessous une approche pour selectionner les po-motifs sur la base de la distribution de l'extension du concept associé.

Un concept formel (X, Y) du treillis cible est décrit par son extension X, qui est un ensemble de paires $(objet, date)$, et son intension Y, qui est représentée par un po-motif. Si les paires contiennent des objets différents, alors elles peuvent être groupées selon les objets. On définit alors $\bar{X}$ qui représente l'ensemble des objets différents dans les paires de X : $\bar{X} = \{o \in O | \exists t \in T, (o,t) \in X\}$, où O est l'ensemble des objets et T l'ensemble des dates.

Definition 2 (Fréquence absolue (ϕ_o)). *La fréquence absolue d'un objet $o \in \bar{X}$, notée ϕ_o, est égale au nombre de paires distinctes de X où o apparaît. On définit $\bar{X}_\phi = \{(o, \phi_o) \, | o \in \bar{X}\}$.*

La figure 6 donne les distributions des objets dans les extensions des concepts `Concept_1` et `Concept_2`, où $\bar{X_1} = \bar{X_2} = \{S1, S2, S3, S4\}$. On a $\bar{X_1}_\phi = \{(S1, 7), (S2, 3), (S3, 7), (S4, 2)\}$ et $\bar{X_2}_\phi = \{(S1, 5), (S2, 4), (S3, 2), (S4, 8)\}$.

Definition 3 (Support et Richesse). *On appelle support de Y le nombre de paires (objet, date) de X. La richesse de Y, notée ρ, est définie comme la cardinalité de $\bar{X}$.*

Definition 4 (Distribution (IQV)). *La distribution de X rend compte du nombre de fois où les objets de $\bar{X}$ apparaissent dans X. Elle est donnée par l'indice de variation qualitative (IQV,*

FIG. 6: Représentation des valeurs ϕ_o pour les objets de deux extensions de concepts.

(Frankfort-Nachmias et Leon-Guerrero, 2010)) qui est fondé sur le rapport entre les valeurs observées dans $\bar{X}_\phi$ et le nombre total des valeurs possibles dans $\bar{X}_\phi$.

$$IQV = \frac{\rho\left(|X|^2 - \sum_{i=1}^{\rho} {\phi_o}_i{}^2\right)}{|X|^2\,(\rho - 1)} \tag{1}$$

Les deux concepts de la figure 6 ont pour support $|X_1| = |X_2| = 19$ et pour richesse $\rho_1 = \rho_2 = 4$. La distribution de `Concept_1` a pour valeur $IQV_1 = \frac{4\left[19^2 - \left(7^2+3^2+7^2+2^2\right)\right]}{19^2(4-1)} = 0.92$ tandis que celle de `Concept_2` vaut $IQV_2 = 0.93$.

Frankfort-Nachmias et Leon-Guerrero (2010) présentent un ensemble de mesures statistiques de la distribution. Notre choix d'utiliser l'indice IQV, qui est recommandé pour calculer une distribution, s'appuie sur l'observation que les objets de $\bar{X}$ n'ont pas d'ordre intrinsèque. L'indice IQV varie de 0 à 1 : si $\bar{X}$ ne contient qu'un objet, il n'y a aucune diversité et IQV vaut 0 ; à l'inverse, si tous les objets de $\bar{X}$ ont le même ϕ_o, la distribution est égale et IQV vaut 1. Ainsi, dans la figure 6 le po-motif du `Concept_2` est plus intéressant car la distribution de l'extension est meilleure. Toutefois, si les po-motifs ont des valeurs différentes de support et de richesse, il faut s'appuyer sur les trois informations (support, richesse et distribution) pour sélectionner les po-motifs intéressants.

6 Expérimentations et discussion

Les expérimentations ont été menées en utilisant l'outil RCAExplore [2]. L'algorithme d'extraction et de sélection des po-motifs a été implanté en Java 8. Les données proviennent du projet Fresqueau qui a permis d'intégrer différentes bases concernant l'état des masses d'eau (Bimonte et al., 2015). Trois jeux de données séquentielles (chacun concernant un paramètre Bio d'une certaine classe de qualité) sont analysés : IBD_{bleu}, $IBD_{rougeOrange}$ et $IBGN_{rouge}$. L'objectif est d'extraire des po-motifs à la fois fréquents, précis et associés à de nombreuses stations. Les jeux de données ont été prétraités et transformés comme décrit en section 3. Le tableau 2 donne quelques statistiques pour les étapes d'analyse relationnelle et d'extraction. L'étape d'analyse relationelle utilise l'algorithme IceBerg (Stumme, 2002), qui produit un treillis de concepts fréquents. Un seuil de 10% a été fixé pour les données Bio uniquement (donc pour le treillis cible). Cette valeur permet de limiter l'analyse aux po-motifs associés à

2. `http://dolques.free.fr/rcaexplore`

beaucoup de stations. Toutefois le nombre de po-motifs extraits est important ce qui justifie d'utiliser les métriques présentées ci-dessus pour sélectionner les po-motifs les plus pertinents.

		ARC				Extraction
Indice	Qualité	Entrée		Sortie		Po-motifs
		Bio	PhC	$\mathcal{L}_{\mathcal{K}bios}$	$\mathcal{L}_{\mathcal{K}phcs}$	
IBD	bleu	1208	1845	11116	130107	891
	rougeOrange	132	264	74615	66821	885
IBGN	rouge	86	197	23405	124414	305

TAB. 2: Résultats sur les données du projet Fresqueau : les colonnes 'Entrée' désignent les nombres d'échantillons Bio et PhC, 'Sortie' le nombre de concepts dans les treillis finaux, 'Po-motifs' le nombre de po-motifs extraits associant PhC et Bio.

Ci-dessous nous présentons les résultats obtenus pour le jeu de données *IBGN*$_{rouge}$ (63 stations). La figure 7 affiche la distribution des valeurs d'IQV et le support. Chaque po-motif est représenté par un cercle dont le diamètre est proportionnel à sa richesse. En appliquant les seuils $\theta_{IQV} = 0.97$ et $\theta_{Support} = 20$, on sélectionne les 24 po-motifs à la fois les mieux distribués sur les stations et les plus fréquents. Les po-motifs peuvent ensuite être triés selon le diamètre des cercles, indicateur du nombre de stations associé à chaque motif et donc de l'étendue géographique du motif.

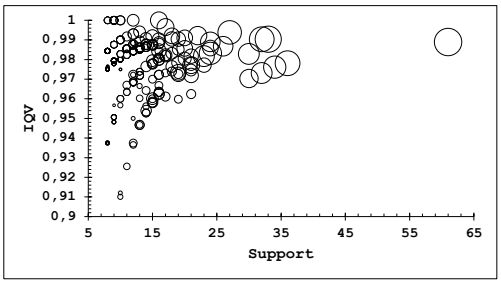


FIG. 7: IQV, support et richesse des po-motifs pour les données *IBGN*$_{rouge}$.

L'interprétation des po-motifs sélectionnés a été réalisée par une hydroécologue. L'ordre de généralisation-spécialisation propre au treillis cible permet à cette experte de commencer son analyse par les motifs les plus fréquents et communs pour aller vers les plus spécifiques ou inversement. La figure 8 est un extrait de la hiérarchie des po-motifs mettant en évidence la relation connue entre `MOOX` (matières organiques) et `IBGN` : $\langle(\texttt{MOOX}_{\texttt{rouge}})\rangle$ couvre 32% des stations étudiées. $\langle(\texttt{PHOS}_{\texttt{rouge}})\rangle$, qui recouvre plus de 33% des stations, est aussi un po-motif intéressant car il montre l'impact possible des pollutions au phosphore sur les macro-invertébrés (`IBGN`), ce qui est moins connu. Par ailleurs, grâce à la structure hiérarchique, l'experte peut identifier facilement et rapidement que la combinaison de `PHOS` *rouge* et `MOOX` *rouge* (le po-motif de `CKbios_3006`) a aussi un impact négatif fort sur les macro-invertébrés.

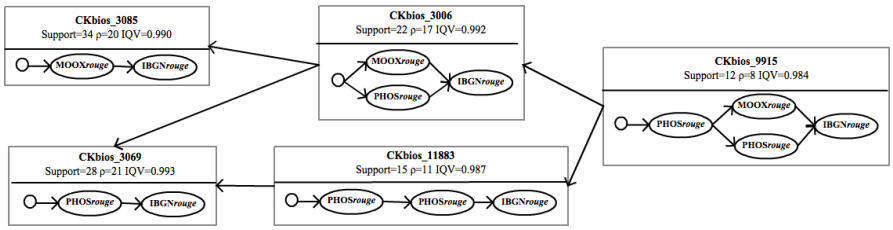


FIG. 8: Extrait de la hiérarchie des po-motifs pour $IBGN_{rouge}$.

7 Conclusion

Dans cet article, nous avons proposé une approche originale pour explorer des données temporelles avec l'ARC. L'approche comporte un processus complet pour explorer des données séquentielles comprenant : 1) une phase d'analyse relationnelle de concepts, s'appuyant sur un modèle temporel des données qui met en exergue les objets d'intérêt, 2) une phase d'extraction de po-motifs à partir des résultats de l'ARC et 3) une phase de sélection des po-motifs intéressants en utilisant des mesures de distribution, richesse et support sur les extensions de concepts. Nous avons évalué notre proposition sur des jeux de données réels issus du projet Fresqueau. Néanmoins, une analyse systématique des po-motifs obtenus reste à faire par un expert du domaine pour confirmer la pertinence de la méthode de sélection proposée.

Dans l'avenir, nous étudierons l'intérêt des attributs relationnels vagues, qui mettent en évidence des liens généraux entre les paramètres biologiques et physico-chimiques. Nous approfondirons l'étude des mesures d'intérêt fondées sur l'extension des concepts, et rechercherons également d'autres méthodes pour réduire le nombre de po-motifs extraits.

Références

Agrawal, R. et R. Srikant (1995). Mining sequential patterns. In *International Conference on Data Engineering*, ICDE, pp. 3–14.

Arévalo, G., J.-R. Falleri, M. Huchard, et C. Nebut (2006). Building abstractions in class models : Formal concept analysis in a model-driven approach. In *MoDELS 2006*, pp. 513–527.

Bimonte, S., K. Boulil, A. Braud, S. Bringay, F. Cernesson, X. Dolques, M. Fabrègue, C. Grac, N. Lalande, F. Le Ber, et M. Teisseire (2015). Un système décisionnel pour l'analyse de la qualité des eaux de rivières. *Ingénierie des Systèmes d'Information 20*(3), 143–167.

Buzmakov, A., E. Egho, N. Jay, S. O. Kuznetsov, A. Napoli, et C. Raïssi (2015). On mining complex sequential data by means of FCA and pattern structures. *ArXiv e-prints*.

Casas-Garriga, G. (2005). Summarizing sequential data with closed partial orders. In *2005 SIAM International Conference on Data Mining*, pp. 380–391.

Fabrègue, M., A. Braud, S. Bringay, C. Grac, F. Le Ber, D. Levet, et M. Teisseire (2014). Discriminant temporal patterns for linking physico-chemistry and biology in hydro-ecosystem assessment. *Ecological Informatics 24*, 210–221.

Fabrègue, M., A. Braud, S. Bringay, F. Le Ber, et M. Teisseire (2015). Mining closed partially ordered patterns, a new optimized algorithm. *Knowledge-Based Systems 79*, 68–79.

Ferré, S. (2007). The efficient computation of complete and concise substring scales with suffix trees. In *Formal Concept Analysis*, pp. 98–113. Springer Berlin Heidelberg.

Frankfort-Nachmias, C. et A. Leon-Guerrero (2010). *Social Statistics for a Diverse Society*, Chapter Measures of Variability. SAGE Publications.

Ganter, B. et R. Wille (1999). *Formal Concept Analysis : Mathematical Foundations*. Springer Berlin Heidelberg.

Klimushkin, M., S. Obiedkov, et C. Roth (2010). Approaches to the selection of relevant concepts in the case of noisy data. In *Formal Concept Analysis*, pp. 255–266. Springer Berlin Heidelberg.

Kuznetsov, S. O. (2007). On stability of a formal concept. *Annals of Mathematics and Artificial Intelligence 49*(1-4), 101–115.

Nica, C., X. Dolques, A. Braud, M. Huchard, et F. Le Ber (2015). Exploration de données temporelles avec des treillis relationnels. In *EGC, Actes de l'atelier GAST*, pp. 45–56.

Pei, J., H. Wang, J. Liu, K. Wang, J. Wang, et P. S. Yu (2006). Discovering frequent closed partial orders from strings. *IEEE Transactions on Knowledge and Data Engineering 18*(11), 1467–1481.

Poelmans, J., P. Elzinga, S. Viaene, et G. Dedene (2010). A Method based on Temporal Concept Analysis for Detecting and Profiling Human Trafficking Suspects. In *Artificial Intelligence and Applications, AIA 2010, Innsbruck, Austria*, pp. 1–9.

Rouane-Hacene, M., M. Huchard, A. Napoli, et P. Valtchev (2013). Relational concept analysis : Mining concept lattices from multi-relational data. *Annals of Mathematics and Artificial Intelligence 67*(1), 81–108.

Stumme, G. (2002). Efficient data mining based on formal concept analysis. In A. Hameurlain, R. Cicchetti, et R. Traunmüller (Eds.), *Database and Expert Systems Applications*, pp. 534–546. Springer Berlin Heidelberg.

Wolff, K. E. (2001). Temporal Concept Analysis. In *ICCS-01 Workshop on Concept Lattice for KDD, 9th International Conference on Conceptual Structures*, pp. 91–107.

Summary

This paper presents a new method of mining temporal data, using Relational Concept Analysis (RCA), that is applied to sequential datasets, dealing with biological and physico-chemical (PhC) parameters sampled in waterbodies. Our aim is to reveal meaningful and hierachical partially ordered patterns (po-patterns) linking the two types of parameters. We propose a comprehensive temporal data mining process starting by using RCA on an ad hoc temporal data model. Then, we continue with the extraction of sets of subsequences summarized as po-patterns. Finally, we select relevant po-patterns, using measures based on the distribution of the concept extents. This process is assessed through some quantitative statistics and qualitative interpretations resulting from experiments carried out on real sequential datasets.