Une Approche d'Extraction de Motifs Graduels (Fermés) Fréquents Sous Contrainte de la Temporalité

Jerry Lonlac*,** Yannick Miras** Aude Beauger**
Marie Pailloux* Jean-Luc Peiry** Engelbert Mephu Nguifo*

*CNRS, UMR 6158, LIMOS, Université Clermont Auvergne, F-63173 Aubière, France
{lonlac, pailloux, mephu}@isima.fr
**CNRS, UMR 6042, GEOLAB, Université Clermont Auvergne, F-63000 Clermont-Ferrand
{yannick.miras, aude.beauger, jean-luc.peiry}@univ-bpclermont.fr

Résumé. La fouille de motifs graduels a pour but la découverte de co-variations fréquentes entre attributs numériques dans une base de données. Plusieurs algorithmes d'extraction automatique de tels motifs ont été proposés. La principale différence entre ces algorithmes réside dans la sémantique de variation considérée. Dans certains domaines d'application, on trouve des bases de données dont les objets sont munis d'une relation d'ordre temporel. Ainsi, du fait de leur sémantique de variation, les algorithmes de la littérature sont inadaptés pour de telles données. Dans ce contexte, nous proposons une approche de fouille de motifs graduels sous contrainte d'ordre temporel, qui réduit le nombre de motifs générés. Une étude expérimentale sur des bases de données paléoécologiques permet d'apprendre les groupements d'indicateurs qui modélisent l'évolution de la biodiversité. Les connaissances apportées par ces groupements montre l'intérêt de notre approche pour le domaine environnemental.

1 Introduction

Les motifs graduels qui capturent les corrélations d'ordre de la forme "plus/moins X, plus/moins Y" jouent un rôle important dans plusieurs applications du monde réel où le volume de données numériques à gérer est important, c'est le cas de données biologiques, ou de données médicales. Les algorithmes de fouille de données sont le plus souvent utilisés pour extraire automatiquement de tels motifs (Berzal et al., 2007; Masseglia et al., 2008; Di-Jorio et al., 2008, 2009; Laurent et al., 2009; Do et al., 2015).

En effet, Berzal et al. (2007) utilisent pour la première fois les méthodes de fouille de données à travers une adaptation de l'algorithme *Apriori* pour extraire les motifs graduels et évaluent le support de ces motifs en considérant tous les couples d'objets possibles. Dans Masseglia et al. (2008), les auteurs introduisent les motifs séquentiels graduels pour rendre compte de la force de modification (accélération). Deux autres méthodes d'extraction sont proposées dans Di-Jorio et al. (2008) et Di-Jorio et al. (2009), la différence entre elles étant liée au mode de calcul du support. En effet, dans Di-Jorio et al. (2008), étant donné un motif graduel, les auteurs proposent une heuristique permettant d'éliminer les objets qui empêchent le maximum

de lignes de la base d'être ordonné. L'ensemble de conflits d'un objet étant constitué de tous les objets qui sont en conflit avec celui-ci. Dans Di-Jorio et al. (2009), une méthode exacte fondée sur l'utilisation de structures binaires est proposée. Dans Laurent et al. (2009), les auteurs proposent un algorithme qui combine les principes de plusieurs approches existantes et bénéficie des propriétés efficaces permettant de calculer le support. En effet, ces derniers considèrent la formulation proposée dans Berzal et al. (2007), et proposent un algorithme qui exploite la structure binaire utilisée dans Di-Jorio et al. (2009). Ils considèrent le coefficient de corrélation de Kendall qui calcule le nombre de paires d'objets ordonnables (concordantes) ou non (discordantes) dans la base de données pour être en accord avec le motif graduel considéré.

La plupart de ces méthodes se heurtent au problème de gestion de la quantité très élevée de motifs extraits. Sur certaines données, le nombre de motifs graduels fréquents peut être important, rendant leur interprétation par l'expert quasiment impossible. Une façon de réduire leur nombre est d'utiliser des représentations condensées de motifs. En effet, à partir d'un ensemble de motifs spécifiques, comme les motifs graduels fermés (Ayouni et al., 2010), il est possible de régénérer l'ensemble de tous les motifs graduels. De plus, les motifs fermés permettent d'éviter d'avoir des informations redondantes. Dans cet ordre d'idée, Do et al. (2015) proposent un algorithme fondé sur le principe de l'algorithme LCM pour fouiller les motifs graduels fermés fréquents en temps linéaire.

D'autre part, les algorithmes proposés dans la littérature pour la fouille de motifs graduels ne supposent aucune contrainte temporelle sur les données. Cependant, il existe des domaines d'application où l'on trouve des bases de données numériques dont les objets ont une signification temporelle, c'est le cas de données paléoécologiques. Étant donné cette contrainte, on peut s'intéresser uniquement aux motifs dont l'ordre des objets concordants respectent l'ordre temporel, les autres motifs n'ayant pas de signification pertinente. De ce fait, les algorithmes proposés dans la littérature sont inadaptés pour l'extraction des motifs graduels dans ces bases paléoécologiques. Nous proposons donc dans cet article, une approche d'extraction de motifs graduels qui prend en considération la contrainte de temporalité entre les objets. Une telle approche est adaptée au contexte et permet de rechercher uniquement des motifs qui apportent des informations utiles et pertinentes répondant aux intérêts de l'utilisateur. Nous définissons, également dans ce même cadre, d'autres contraintes que nous utilisons en post-traitement pour réduire le nombre de motifs extraits.

Cet article est organisé comme suit : nous présentons en section 2 un contexte où l'on trouve des données numériques avec des problématiques de fouille de motifs graduels sous contrainte d'ordre temporel. Après avoir introduit la notion de motifs graduels, notre approche d'extraction de motifs graduels fermés fréquents sous contrainte temporelle est décrite en section 4. Avant de conclure, les résultats des expérimentations menées sur des données paléoécologiques, et leur interprétation sont présentés.

2 Domaine d'application : la paléoécologie

Les recherches paléoécologiques permettent de reconstruire au cours du temps les dynamiques écologiques et l'évolution de la biodiversité (par exemple l'évolution de la végétation ou du fonctionnement d'un écosystème lacustre) sous l'influence des variations climatiques et des activités humaines (par exemple agriculture et pastoralisme) (Smol et al., 2001). La reconstruction de ces trajectoires écologiques sur 7 millénaires pour le lac d'Aydat, situé dans

la Chaîne des Puys en région Auvergne Rhône-Alpes et menacé d'eutrophisation, permet de mieux connaître son état écologique actuel et de concourir à l'établissement de gouvernances viables (Miras et al., 2015). Cette recherche est fondée, sur l'abondance de différents indicateurs paléoécologiques (grains de pollen et spores de végétaux ; micro-fossiles non polliniques : différentes formes de résistance du phytoplancton et du zooplancton ; diatomées) conservés dans l'enregistrement sédimentaire lacustre. Toutes ces données paléoécologiques sont ensuite stockées dans des bases de données numériques.

Les données paléoécologiques sont constituées d'un ensemble d'attributs à valeurs numériques correspondant à la quantité de chaque indicateur paléoécologique contenu dans un enregistrement sédimentaire prélevé, par des opérations de carottage, au sein d'un écosystème lacustre. La séquence sédimentaire obtenue est ensuite datée, échantillonnée, et pour chaque échantillon, à une profondeur donnée, une date est calculée. L'abondance de chaque indicateur est ensuite relevée pour chaque échantillon. Les objets de cette base de données correspondent aux différentes dates obtenues sur l'enregistrement considérée, et les colonnes aux différents indicateurs paléoécologiques relevés.

Plus formellement, soient $\mathcal{D}$ les dates des différents échantillons de la séquence sédimentaire, $\mathcal{I}$ les différents indicateurs paléoécologiques relevés à ces dates $\mathcal{D}$, alors le tableau de données paléoécologiques est défini par $\Delta = \mathcal{D} \times \mathcal{I}$. Les notations suivantes caractérisent les données contenues dans Δ.

Soit $d \in \mathcal{D}$ et $i \in \mathcal{I}$:

— $\Delta(d, i)$ indique le nombre d'instances de l'indicateur i présents à la date d ;

— $\Delta(d, i) = 0$ indique l'absence de l'indicateur i à la date d ;

— $\Delta(\mathcal{D}, i)$ montre l'évolution temporelle de l'indicateur i ;

— $\Delta(d, \mathcal{I})$ caractérise l'état des conditions paléoécologiques à la date d.

Une particularité des données paléoécologiques contenues dans Δ est qu'elles sont évolutives. De plus, le tableau Δ comportent généralement peu de lignes (les dates des échantillons) au regard du nombre de colonnes (nombre d'indicateurs paléoécologiques) et est très peu dense (il contient un grand nombre de valeurs nulles).

Le tableau 1 est un extrait de données paléoécologiques, ces colonnes sont étiquetées par les noms scientifiques des différents indicateurs. Pour des raisons de simplicité, afin d'illustrer notre approche dans la suite, nous présentons ici une base restreinte à 7 indicateurs et 7 dates. Nous considérerons le tableau 1 que nous désignons par Δ, comme une base de transactions contenant des attributs numériques. Ainsi, les 7 indicateurs paléoécologiques correspondront aux différents attributs de notre base de données et les 7 dates identifieront les différents objets.

	Poaceae	**Secale.t**	**Rumex.ace**	**Equisetum**	**Plantago.l**	**Filipendula.v**	**Coprofilous.f**
d_1	84	61	7	0	1	2	0
d_2	116	36	4	1	11	2	31
$d3$	90	52	2	3	5	2	13
d_4	124	34	1	5	12	1	36
d_5	102	49	0	6	7	0	17
d_6	135	17	0	1	18	0	62
d_7	106	40	3	1	9	0	18

TAB. 1 – *Exemple de base de données paléoécologiques* Δ.

Il est alors question pour les experts de la recherche paléoécologique, de les analyser afin d'en extraire des connaissances spécifiques telles que la mise au jour des groupements de co-

évolution d'indicateurs multi-variés paléoécologiques (par exemple des groupements constitués de grains de pollen, de micro-fossiles non polliniques et de diatomées) nécessaires à la compréhension de l'évolution de la biodiversité et du fonctionnement d'un écosystème au cours du temps. Toutes ces informations implicites sont généralement recherchées par les experts en paléoenvironnement en utilisant les méthodes d'analyses statistiques classiques. Celles-ci reposent le plus souvent sur un simple tracé d'un graphique contenant des courbes d'évolution des différents indicateurs paléoécologiques à partir des données paléoécologiques (voir la figure située à l'adresse *http ://mobipaleo.univ-bpclermont.fr/*, sur le lien "$Visualiser\ l'evolution\ des\ especes\ ici$") (figure 1) et sur une comparaison empirique de ces courbes afin de relever des groupes de courbes qui évoluent à des périodes identiques.

La figure 1 montre un graphique contenant les courbes d'évolution de 178 indicateurs paléoécologiques sur 110 dates. Chaque indicateur est représenté sur la légende du graphique par son nom scientifique. Il apparaît évident qu'il est très fastidieux pour les experts de la paléoécologie d'identifier de manière empirique les différentes coévolutions de ces indicateurs à partir de ce graphique. De plus le risque est d'exclure un traitement exhaustif des données et de laisser échapper éventuellement la pépite d'information pertinente : les groupes à faible coévolution mais à fort impact sur le changement de la biodiversité. Dans ce contexte, les motifs graduels s'avèrent adaptés pour résoudre le problème d'extraction automatique des groupements de coévolution d'indicateurs multi-variés paléoécologiques.

Dans ce papier, nous montrons que, sans la prise en compte de la contrainte de temporalité entre les objets de la base de données, les sémantiques proposées dans les algorithmes de fouille de motifs graduels de la littérature ne seront pas adaptés au contexte de la temporalité (exemple des bases de données paléoécologiques). Pour ces données, cette contrainte permet d'éviter au cours du processus de fouille, de générer des motifs graduels qui ne correspondent pas aux coévolutions, ce qui permet des gains de temps de calcul et de mémoire consommée considérable. L'approche proposée apporte de l'information supplémentaire par rapport aux techniques classiques utilisées jusqu'à présent pour l'analyse de données paléoéologiques, qui sont essentiellement des méthodes statistiques et de classification.

3 Les motifs graduels

Dans cette section, nous rappelons la notion de motifs graduels et l'illustrons sur une base de données paléoécologiques.

Si on considère une base de données numériques Δ contenant un ensemble d'objets $\mathcal{D} = \{ d_1, ..., d_n\}$ décrit par un ensemble d'attributs $\mathcal{I} = \{ i_1, ..., i_m\}$, les motifs graduels extraits de Δ sont de la forme "plus/moins i_1, . . ., plus/moins i_k" $(k \leq m)$. Ces motifs sont définis sur un sous-ensemble de $\mathcal{D}$ dont les éléments sont associés à un ordre croissant ou décroissant. Nous désignons par $d_j[i_k]$ la valeur de l'attribut i_k sur l'objet d_j.

Définition 1 (item graduel) *Soit Δ une base de données définie sur un ensemble d'attributs numériques $\mathcal{I}$, un item graduel est défini sous la forme i^*, où i est un attribut de $\mathcal{I}$ et $* \in \{+, -\}$ et le domaine des valeurs de i est muni d'une relation d'ordre total. Lorsque $*$ correspond à '+', cela signifie que la valeur de i augmente, et lorsque $*$ correspond à '$-$' cela signifie que la valeur de i diminue.*

Un itemset (motif) graduel $s = (i_1^{*1}, ..., i_k^{*k})$ est un ensemble non vide d'items graduels.

Pour l'illustration, nous considérons la base de données du tableau 1. Elle est constituée d'un ensemble d'objets (les différentes dates de d_1 à d_7) et d'un ensemble d'attributs (les noms scientifiques des différents indicateurs paléoécologiques). Cette base indique la quantité de chaque indicateur à chaque date. Relativement à cette base, $(Poaceae^+)$ est un item graduel, tandis que $\{Poaceae^+, Secale.t^-\}$ est un motif graduel qui indique que *"plus le nombre d'indicateur Poaceae croît , plus le nombre d'indicateur Secale.t décroît"*.

Définition 2 (motif graduel complémentaire) *Soit $s = (i_1^{*1}, ..., i_k^{*k})$ un itemset graduel, et c une fonction telle $c(+) = " - "$ et $c(-) = " + "$ alors $c(s)$ désigne le complémentaire de s.*

Le complémentaire d'un motif graduel est encore appelé *motif symétrique*. Le symétrique du motif graduel $\{Poaceae^+, Secale.t^-\}$ est $\{Poaceae^-, Secale.t^+\}$.

Le calcul du support d'un motif graduel dans une base de données Δ revient à mesurer à quel point le motif est présent dans Δ. Un motif graduel est dit fréquent si son support (fréquence) est supérieur ou égal à un seuil minimal fixé par l'utilisateur. La problématique d'extraction de motifs graduels fréquents consiste à trouver l'ensemble de tous les motifs graduels fréquents dans une base de données numériques, selon un seuil de fréquence minimal.

Plusieurs approches d'extraction automatique de motifs graduels ont été proposées dans la littérature. La principale différence entre ces approches réside dans leur mode de calcul du support et dans la sémantique de variation considérée. Ces approches ne supposent aucune contrainte de temporalité entre les objets, ce qui est inadapté dans le cas des bases de données numériques dont les objets sont munis d'une contrainte d'ordre temporel (exemple de la base de données Δ). Un autre problème non pris en considération par la plupart des approches de la littérature est le problème des valeurs égales. Ce problème dont l'intérêt a été souligné pour la première fois par Do et al. (2015) n'a pas retenu beaucoup d'attention et se retrouve dans plusieurs bases de données réelles. En effet, dans certaines bases de données numériques, l'on trouve des objets avec des valeurs identiques pour un même attribut (exemple de l'attribut *Filipendula.v* de la base de données Δ). Ainsi, Do et al. (2015) proposent une solution pour ce problème, mais uniquement lors de l'évaluation du support du motif graduel.

Dans ce travail, nous proposons une approche pour extraire automatiquement les motifs graduels (fermés) fréquents dans des bases de données numériques qui prend en compte la contrainte de temporalité entre les objets de la base de données au cours du processus de fouille tout en s'affranchissant du problème des valeurs égales. Elle est fondée sur le principe de la formulation donnée par Berzal et al. (2007) et permet de réduire le temps d'exécution et la mémoire en éliminant au cours du processus de fouille les motifs inutiles et inintéressants. L'application de notre approche sur des données paléoécologiques permet d'extraire les motifs graduels (fermés) fréquents qui correspondent aux groupements d'indicateurs multivariés permettant de modéliser l'évolution des écosystèmes considérés.

4 Fouille de motifs graduels sous contrainte temporelle

Nous présentons dans cette section, notre processus d'extraction de motifs graduels (fermés) fréquents dont la séquence d'objets concordants respecte l'ordre temporel, et qui s'affranchit du problème de valeurs égales.

4.1 Gradualité sous contrainte de la temporalité

La sémantique de temporalité que nous proposons de prendre en compte ici pour notre contexte d'étude se démarque de celle proposée par Berzal et al. (2007) dans la mesure où, d'une part, elle évite de rechercher les motifs graduels inutiles (motifs dont la séquence d'objets concordants ne respectent pas l'ordre temporel). D'autre part, les motifs que nous recherchons ne sont pas symétriques comme les motifs obtenus dans Berzal et al. (2007) (que nous appellerons ici motifs graduels classiques). En effet, dans Berzal et al. (2007), les auteurs proposent de construire à partir d'une base de données numériques Δ, une nouvelle base de données Δ', constituée de toutes les paires d'objets et d'utiliser l'algorithme *Apriori* sur Δ' pour rechercher tous les itemsets fréquents, lesquels vont correspondre aux motifs graduels fréquents de Δ. Ces auteurs ne considèrent que deux items dans Δ' : à savoir $i^*(* \in \{\geq, \leq\})$.

Dans le souci de gérer le problème de valeurs égales, nous construisons à partir de la base de données numériques Δ initiale, une nouvelle base de données Δ' contenant trois items au lieu de deux : à savoir $i^*(* \in \{<, >, o\})$, où l'item i^o indique les cas où les valeurs d'un attribut i restent constantes. Dans la suite de ce papier, l'opérateur "<" (respectivement ">") sera désigné par "-" (respectivement "+"), et l'opérateur d'égalité sera désigné par le symbole "o". Plus formellement, la nouvelle base Δ' est définie comme suit :

Définition 3 *Soient $\mathcal{I} = \{i_1, .., i_m\}$ un ensemble d'attributs numériques et $\mathcal{D} = \{d_1, .., d_n\}$ un ensemble d'objets **"ordonné"** où chaque objet d_j avec $j \in [1, n]$ contient une valeur numérique pour chaque attribut dans $\mathcal{I}$. Le problème de fouille de motifs graduels fréquents dans la base $\Delta = \mathcal{D} \times \mathcal{I}$ peut être ramené à un problème de fouille d'itemsets fréquents dans une nouvelle base de données contenant des attributs catégoriels $\Delta' = \mathcal{D}' \times \mathcal{I}'$ telle que :*

- $\mathcal{D}' = \{d'_1, .., d'_{n-1}\}$, $|\mathcal{D}'| = n - 1$
- $\mathcal{I}' = \{i_1^+, i_1^-, i_1^o, .., i_m^+, i_m^-, i_m^o\}$, $|\mathcal{I}'| = 3 \times |\mathcal{I}|$
- $\forall d'_j \in \mathcal{D}'$, $i_k^{*p} \in d'_j \Leftrightarrow d_{j+1}[i_k] *p\ d_j[i_k]$, *avec* $j \in [1, n-1]$, $k \in [1, m]$, $*p \in \{+, -, o\}$.

Définition 4 (motif graduel respectant la temporalité) *Soient $\Delta = \mathcal{D} \times \mathcal{I}$ une base de données numériques, s un motif graduel fréquent extrait de Δ et $L_s = < d_{l_1}, ..., d_{l_j} >$ la séquence d'objets concordants de s. Le motif s respecte l'ordre temporel des objets de Δ si on a l'inégalité suivante : $d_{l_1} < d_{l_2} < ... < d_{l_j}$.*

En utilisant la définition 3, nous construisons à partir de la base de données numériques Δ (voir tableau 1), une nouvelle base de données Δ' (voir tableau 2). Les itemsets fréquents extraits de Δ' correspondent aux motifs graduels fréquents de Δ respectant la temporalité.

	Poaceae	Secale.t	Rumex.ace	Equisetum	Plantago.l	Filipendula.v	Coprofilous.f
d'_1	+	-	-	+	+	o	+
d'_2	-	+	-	+	-	o	-
d'_3	+	-	-	+	+	-	+
d'_4	-	+	-	+	-	-	-
d'_5	+	-	o	-	+	o	+
d'_6	-	+	+	o	-	o	-

TAB. 2 – *Base de données Δ' obtenue à partir de données numériques Δ.*

Il faut noter que pour ce travail, nous simplifions notre problème en regardant uniquement les variations entre les objets consécutifs. Cette simplification permet de réduire le temps de construction de la base de données Δ' qui contiendra un plus petit nombre d'objets. Un algorithme d'extraction d'itemsets fréquents classique peut alors être ensuite appliqué sur toute la base Δ' et non sur quelques attributs comme dans Berzal et al. (2007).

Une autre remarque importante est que, en prenant en compte la contrainte de temporalité entre les objets, le support du complémentaire d'un motif graduel ne peut se déduire automatiquement du support de ce motif comme dans le cas des approches de la littérature. On peut formaliser cette différence par le lemme suivant :

Lemme 1 (itemset graduel complémentaire respectant la temporalité) *Soit s, un motif graduel fréquent respectant la contrainte de temporalité tel que $c_t(s)$ est le complémentaire de s. Soit L_s (respectivement $L_{c_t(s)}$) la liste des objets correspondant au motif graduel s (respectivement $c_t(s)$), on a $L_s \cap L_{c_t(s)} = \emptyset$.*

Le lemme 1 indique que, dans le contexte de la temporalité, la génération de la moitié des motifs graduels fréquents n'est pas suffisante pour déduire automatiquement l'autre moitié comme dans les approches de la littérature. En effet, les motifs graduels fréquents sont symétriques si on ne prend pas en compte la contrainte de temporalité entre objets (si s est un motif graduel fréquent alors $c(s)$ est aussi un motif graduel fréquent). Les motifs graduels recherchés dans notre contexte n'étant pas symétriques, il est indispensable de calculer, en plus des supports de tous les motifs graduels, les supports de leur motif complémentaire correspondant.

Proposition 1 *Soient Δ une base de données numériques, C (respectivement C_t) l'ensemble de tous les motifs graduels classiques (respectivement l'ensemble de tous les motifs graduels respectant la contrainte de temporalité) extraits de Δ, nous avons $|C| \geq |C_t|$.*

Motifs graduels fermés : cas du contexte de la temporalité

Les itemsets fermés sont des clés pour obtenir une représentation condensée des motifs sans perte d'information (Pasquier et al., 1999). Un itemset I est dit fermé s'il n'existe aucun itemset I' tel que $I \subset I'$ et $support(I) = support(I')$. Cette notion de fermeture a été introduite pour la première fois dans les motifs graduels dans Ayouni et al. (2010) où les auteurs proposent une paire de fonctions (f, g) définissant un opérateur de fermeture pour les itemsets graduels. Étant donné un ensemble de séquence de transactions $\mathcal{L}$ d'une base de données, f retourne l'itemset graduel P respectant toutes les séquences de transactions dans $\mathcal{L}$ tandis que g retourne l'ensemble des séquences de transactions maximales $\mathcal{L}$ qui respectent les variations de tous les items graduels dans P. Avec ces fonctions, un motif graduel P est dit fermé si $f(g(P)) = P$. Dans Ayouni et al. (2010), les auteurs utilisent ces définitions plutôt dans une étape de post-traitement des motifs. Dans Do et al. (2015), ces définitions sont incluses dans le processus d'extraction de motifs graduels et permet de réduire les temps d'exécution et la consommation mémoire.

Dans notre contexte, les motifs graduels fermés qui peuvent être extraits de la base de données numériques initiale, correspondent aux itemsets fermés extraits de la nouvelle base de données contenant des attributs catégoriels, obtenue à partir de la base de données initiale. Par exemple, les itemsets fermés extraits de la base de données Δ' précédente correspondent aux motifs graduels fermés de la base de données numériques Δ.

Algorithme 1 : $T - GPatterns$

Données : Δ : base de données numériques, $minSupp$: un seuil de support minimal.
Résultat : Γ : motifs graduels (fermés) fréquents.

```
Début
    Δ' ⟵ NumVersCat(Δ) ;
    Γ ⟵ ChercherCoevolution(APRIORI(Δ', minSupp)) ;
    retourner Γ ;
Fin
```

La procédure $NumVersCat$ construit une nouvelle base de données catégorielles Δ' à partir de la base de données numériques initiale Δ en utilisant la définition 3. $APRIORI$ est la procédure utilisée dans Agrawal et Srikant (1994), nous l'utilisons pour générer à partir de Δ', les itemsets fermés fréquents qui correspondent aux motifs graduels fermés fréquents de la base de données Δ. $APRIORI$ étant utilisée par Berzal et al. (2007), nous choisissons de l'utiliser dans l'optique d'une comparaison de notre approche. La procédure $ChercherCoevolution$ recherche les motifs de la forme $\{i_1^{*1}, ..., i_k^{*k}\}$ avec $*p \in \{+, -\}$, $p \in [1, k]$.

4.2 Contraintes pour post-traitement : cas des données paléoécologiques

Nous définissons dans cette section des contraintes adaptées à notre problématique afin de réduire le nombre de motifs trouvés en sélectionnant les plus pertinents.

Étant donné que l'on recherche les coévolutions d'attributs, nous nous intéressons uniquement aux motifs de la forme $\{i_1^{*1}, ..., i_k^{*k}\}$ avec $*p \in \{+, -\}$, $p \in [1, k]$. Par ailleurs, l'augmentation du nombre d'indicateurs paléoécologiques est plus indicative que sa diminution - car différents facteurs peuvent modifier leur représentativité et les groupements de coévolution extraits devant permettre aux experts d'évaluer les impacts à long terme des activités humaines sur la biodiversité, nous proposons d'utiliser en post-traitement les contraintes suivantes :

- C_1 : un groupement d'évolution doit contenir au moins 1 indicateur paléoécologique direct (indicateur primaire) d'activités humaines. Nous avons choisi pour cela l'abondance de grains de pollen de seigle dénommé suivant le nom scientifique *Secale.type*, et marqueur d'activités agricoles (Behre, 1981), ainsi que l'abondance de *spores de champignon coprophile* indicateur d'activités pastorales (Van-Geel, 2001)
- C_2 : un groupement d'évolution doit contenir au moins 1 indicateur direct d'activités humaines avec au moins 1 indicateur sécondaire d'impact anthropique (Behre, 1981). 6 taxons ont été choisis : *Plantago.sp, Plantago.lanceolata, Artemisia, Rumex.acetosella.type, Rannunculus.acris, Poaceae*.
- C_3 : l'unicité de la coévolution d'un groupement d'attributs. Plus formellement, étant donné un ensemble E de motifs graduels fréquents extrait, un motif graduel $s = \{i_1^{*1}, ..., i_k^{*k}\}$ de E est un groupement de coévolution unique si et seulement si :
  - s est soit de la forme $\{i_1^+, ..., i_k^+\}$, ou de la forme $\{i_1^-, ..., i_k^-\}$
  - $\nexists s' = \{i_1^{*1}, ..., i_k^{*k}\}$ tel que $s \neq s'$ et $s \neq c(s')$.

4.3 Expérimentations

Les tests ont été effectués sur trois bases de données numériques d'indicateurs paléoécologiques provenant du lac d'Aydat. La première base utilisée contient 111 objets correspondant à différentes dates identifiées sur l'enregistrement lacustre considéré, et 87 attributs correspondant à différents indicateurs d'anthropisation paléoécologiques (grain de pollen). La deuxième base de données contient 57 objets et 178 attributs (diatomées) liés aux conditions paléohydrologiques (statut trophique de l'eau). La troisième base contient 57 objets et des attributs multi-variés (grains de pollen et diatomées). Nous présentons ensuite quelques motifs graduels fermés fréquents intéressants, extraits de ces différentes bases de données en utilisant *l'algorithme 1*. Ces motifs interprétés et validés par les experts correspondent à des groupements intéressants d'indicateurs d'évolution des hydrosystèmes. Les nouvelles connaissances que révèlent ces groupements dans le domaine de la recherche environnementale montrent l'intérêt de l'approche proposée.

4.3.1 Résultats

Dans nos expérimentations, le seuil minimal de support *(minSupp)* est fixé à 10%. Le choix d'un tel seuil est dû, d'une part, à la faible densité de nos bases données et, d'autre part, ce seuil nous permet d'éviter d'obtenir uniquement des groupements d'indicateurs de la forme $\{i_1^o, ..., i_k^o\}$. Il permet également de conserver le maximum de groupements et d'utiliser en post-traitement les contraintes décrites dans la section 4.2 pour réduire le nombre de groupements obtenus. L'algorithme 1 permet de découvrir dans la première base de données paléoécologiques 2366 motifs graduels fermés fréquents. L'application des contraintes C_1, C_2 et C_3 permet de réduire considérablement le nombre de motifs à 49. Dans la deuxième (respectivement troisième) base de données, 777 (respectivement 284) motifs sont extraits, la contrainte C_3 permet de réduire le nombre de motifs à 296 (respectivement 104). Le tableau 3 présente quelques motifs graduels fermés fréquents parmi les plus pertinents extraits de nos trois bases de données paléoécologiques. Ils correspondent à des groupements d'indicateurs paléoécologiques d'évolution de la biodiversité (floristique et limnologique) au cours du temps.

4.3.2 Interprétation des résultats

Les motifs extraits sont pertinents car ils sont indicateurs soit d'impacts anthropiques (motifs de 1 à 22 du tableau 3), soit d'enrichissement trophique (motifs de 23 à 43).

Ces motifs comprennent des indicateurs en coévolution qui sont cohérents dans la mesure où les indicateurs traduisent des conditions paléoécologiques similaires. Par exemple 86% des motifs contenant un indicateur direct contiennent au moins un indicateur pollinique secondaire d'impact anthropique. Par ailleurs 100% des motifs fondés sur les diatomées traduisent une coévolution de taxons (indicateurs paléoécologiques) indiquant un statut trophique élevé des eaux du lac d'Aydat. Ces motifs constituent donc bien des groupements fonctionnels d'indicateurs d'activités humaines ou d'état paléoécologique. Ils permettent également de renforcer, voire de préciser le potentiel paléoécologique de certains taxons.

— Cette étude permet de suggérer une possible discrimination des indicateurs polliniques secondaires d'impact anthropique en rattachant :

  - certaines coévolutions plutôt à l'agriculture : *Secale.type* coévoluant de manière significative avec *Plantago.lanceolata* dans les motifs de 9 à 15 et 17 ou avec *Ranunculus.acris* dans les motifs de 4 à 8 et 15.
  - et d'autres coévolutions plutôt avec l'activité pastorale : *Coprofilous.Fungi coévoluant de manière répétée avec Plantago.sp* et *Rumex.acetosella.type* dans les motifs 1 à 3 et 22.
- Cette étude, pour ce qui concerne les motifs fondés sur les diatomées apporte des informations sur les preferenda écologiques des taxons. Par exemple, alors que peu d'informations existent pour le taxon *ECPM* (*Encyonopsis minuta*) dans la littérature, nos motifs (les motifs 23 et 24) associent clairement ce taxon avec des espèces eutrophes (par exemple *SCON, SRPI*). A l'inverse, nos motifs (les motifs 28 et 30) confèrent un statut bien plus ubiquiste au taxon *PTLA* dont les préférences écologiques semblent davantage s'étaler entre des conditions mésotrophes et hyper-eutrophes respectivement indiquées par les taxons *AUSU, FNAN*, d'une part, et par *SHAN, STMI* d'autre part.

Enfin, ces motifs permettent d'interroger ou de valider le potentiel paléoécologique de certains taxons. En effet, certains motifs font coévoluer deux marqueurs classiques du pastoralisme (*Plantago.sp et Coprofilous.Fungi, (Behre, 1981; Van-Geel, 2001)*) avec *Polygonum* (les motifs 1 et 2) ou les *Rosaceae* (motif 3) qui ne sont pas ou peu associés à des activités humaines en paléoécologie. Une situation similaire est retrouvée dans les motifs 4 de 7, 11 à 15, et dans les motifs 18, 19 et 21 où certains taxons comme *Filipendula.vulgaris, Apiaceae, Equisetum* coévoluent avec des indicateurs classiques d'impacts anthropiques. Ceci suggère le caractère local des activités humaines retracées et l'impact de celles-ci sur l'écosystème.

Ces motifs extraits confirment par ailleurs l'utilité de certains micro-fossiles non polliniques dans la caractérisation d'un enrichissement trophique d'un lac (Miras et al., 2015). C'est particulièrement le cas des œufs de rotifères tels *Conochilus.natans.type, Trichocerca.cylindrica, Anuraeopsis.fissa.type* dans les motifs de 33 à 35, 37 à 38 et 42 ou de certaines formes algales comme *Hdv-128* (Van-Geel, 2001) dans le motif 40. Certains motifs constituent même des groupements d'indicateurs paléoécologiques multi-variés regroupant des espèces ou des morphotypes franchement eutrophes, voir hyper-eutrophes vivant dans des eaux riches en matière organique fermentescible (motif 39).

5 Conclusion et perspectives

Dans ce papier, nous proposons une approche de fouille de motifs graduels fermés fréquents dans des bases de données numériques dont les objets sont munis d'une relation d'ordre temporel. Nous avons présenté un domaine (la Paléoécologie) où l'on trouve des données avec des problématiques de fouille de motifs graduels sous contrainte de temporalité. Nous montrons que, dans ce contexte, la prise en compte de la contrainte de temporalité au cours du processus de fouille permet de réduire significativement le nombre de motifs en éliminant ceux dont la séquence d'objets concordants ne respecte pas la temporalité. Un algorithme dédié à l'extraction automatique des motifs graduels fermés fréquents dans ce contexte de temporalité a été proposé. Les expérimentations menées sur des données paléoécologiques ont permis d'appréhender des groupements fonctionnels de coévolution d'indicateurs paléoécologiques qui modélisent l'évolution de la biodiversité au cours du temps. Les connaissances apportées par ces groupements montrent l'intérêt de l'approche pour la Paléoécologie. L'objectif ici était

de montrer comment la contrainte de temporalité peut être prise en compte au cours du processus de fouille de motifs graduels. Nous avons proposé une première solution qui simplifie le problème en considérant les gradualités uniquement entre les objets consécutifs. Il serait intéressant de considérer dans ce contexte de temporalité, d'autres sémantiques de gradualité de la littérature. Ce dernier point fait l'objet des travaux en cours.

Remerciements. Ce travail est soutenu par la région Auvergne Rhône-Alpes et l'Union Européenne dans le cadre du projet *MobiPaléo (http ://mobipaleo.univ-bpclermont.fr/)* du CPER 2014. Nous remercions les relecteurs anonymes pour leurs remarques constructives.

Références

Agrawal, R. et R. Srikant (1994). Fast algorithms for mining association rules in large databases. In *VLDB, Santiago de Chile, Chile, September 12-15*, pp. 487–499.

Ayouni, S., A. Laurent, S. B. Yahia, et P. Poncelet (2010). Mining closed gradual patterns. In *Artificial Intelligence and Soft Computing, 10th International Conference, ICAISC, Zakopane, Poland, June 13-17, Part I*, pp. 267–274.

Behre, K. (1981). The interpretation of anthropogenic indicators in pollen diagrams. In *Pollen et Spores : 23*, pp. 225–245.

Berzal, F., J. C. Cubero, D. Sánchez, M. A. V. Miranda, et J. Serrano (2007). An alternative approach to discover gradual dependencies. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 15*(5), 559–570.

Di-Jorio, L., A. Laurent, et M. Teisseire (2008). Fast extraction of gradual association rules: a heuristic based method. In *CSTST, Cergy-Pontoise, France, October 28-31*, pp. 205–210.

Di-Jorio, L., A. Laurent, et M. Teisseire (2009). Mining frequent gradual itemsets from large databases. In *IDA, Lyon, France, August 31 - September 2*, pp. 297–308.

Do, T. D. T., A. Termier, A. Laurent, B. Négrevergne, B. O. Tehrani, et S. Amer-Yahia (2015). PGLCM: efficient parallel mining of closed frequent gradual itemsets. *Knowl. Inf. Syst. 43*(3), 497–527.

Laurent, A., M. Lesot, et M. Rifqi (2009). GRAANK: exploiting rank correlations for extracting gradual itemsets. In *FQAS, Roskilde, Denmark, October 26-28*, pp. 382–393.

Masseglia, F., A. Laurent, et M. Teisseire (2008). Gradual trends in fuzzy sequential patterns. In *IPMU*, pp. 456–463.

Miras, Y., A. Beauger, M. Lavrieux, V. Berthon, K. Serieyssol, V. Andrieu-Ponel, et P. Ledger (2015). Tracking long-term human impacts on landscape, vegetal biodiversity and water quality in the lake aydat (auvergne, france) using pollen, non-pollen palynomorphs and diatom assemblages. In *Palaeogeography, Palaeoclimatology, Palaeoecology*, pp. 76–90.

Pasquier, N., Y. Bastide, R. Taouil, et L. Lakhal (1999). Discovering frequent closed itemsets for association rules. In *ICDT, Jerusalem, Israel, January 10-12*, pp. 398–416.

Smol, H., H. Birks, et W. Last (2001). Tracking environmental change using lake sediments. In *Terrestrial, Algal and Silicaceaous Indicators III. Kluwer Academic Publishers, Dordrecht*, pp. 319–349.

Van-Geel, B. (2001). Tracking environmental change using lake sediments. In *Terrestrial, Algal and Silicaceaous Indicators III. Kluwer Academic Publishers, Dordrecht*, pp. 99–119.

	Motifs
1	Plantago.sp=+, Polygonum=+, Coprofilous.Fungi=+
2	Plantago.sp=-, Polygonum=-, Coprofilous.Fungi=-
3	Rumex.acetosella.type=+, Rosaceae=+, Coprofilous.Fungi=+
4	Secale.type=+, Ranunculus.acris.t=+, Filipendula.vulgaris=+
5	Secale.type=-,Ranunculus.acris.t=-,Filipendula.vulgaris=-
6	Secale.type=+,Ranunculus.acris.t=+,Apiaceae=+
7	Secale.type=+, Ranunculus.acris.t=+, Equisetum=+
8	Secale.type=+, Ranunculus.acris.t=+, Spore.trilete=+
9	Secale.type=+, Plantago.lanceolata=+, Ranunculus.acris.t=+
10	Secale.type=-, Plantago.lanceolata=-, Ranunculus.acris.t=-
11	Secale.type=+, Plantago.lanceolata=+, Apiaceae=+
12	Secale.type=+, Plantago.lanceolata=+, Filipendula.vulgaris=+
13	Secale.type=-, Plantago.lanceolata=-, Filipendula.vulgaris=-
14	Secale.type=+, Plantago.lanceolata=+, Equisetum=+
15	Secale.type=+, Plantago.lanceolata=+, Ranunculus.acris.t=+, Filipendula.vulgaris=+
16	Poaceae=+, Secale.type=+, Ranunculus.acris.t=+
17	Poaceae=+,Secale.type=+,Plantago.lanceolata=+
18	Poaceae=+,Secale.type=+,Apiaceae=+
19	Poaceae=+,Secale.type=+,Equisetum=+
20	Poaceae=+,Secale.type=+,Spore.trilete=+
21	Secale.type=+,Apiaceae=+,Filipendula.vulgaris=+
22	Poaceae=+,Rumex.acetosella.type=+,Coprofilous.Fungi=+
23	ECPM=+,SCON=+,SRPI=+
24	ECPM=+,MCIR=+,SCON=+
25	SCON=+,UDAN=+,EOMI=+
26	SCON=+,PSBR=+,EOMI=+
27	SCON=+,SPAV=+,EOMI=+
28	PTLA=+,SHAN=+,STMI=+
29	AUSU=+,ESLE=+,SPAV=+
30	AUSU=+,FNAN=+,PTLA=+
31	DPST=+,FGRA=+,UUAC=+
32	PLFR=+,STMI=+,Botryococcus=+
33	SSVE=+,Botryococcus=+,Conochilus.natans.type=+
34	Botryococcus=+,Trichocerca.cylindrica.type=+,Anuraeopsis.fissa.type=+
35	PSBR=+,EOMI=+,Trichocerca.cylindrica.type=+
36	PSBR=+,EOMI=+,Spirogyra=+
37	EOMI=+,Spirogyra=+,Trichocerca.cylindrica.type=+
38	Pediastrum=+,Botryococcus=+,Conochilus.natans.type=+
39	SHAN=+,SMED=+,Botryococcus=+
40	SPAV=+,Botryococcus=+,HdV.128=+
41	SPAV=+,Botryococcus=+,Spirogyra=+
42	SSVE=+,Pediastrum=+,Botryococcus=+,Conochilus.hippocrepis.type=+
43	AFOR=+,SPAV=+,Pediastrum=+,Botryococcus=+

TABLE 3 – *Motifs graduels intéressants extraits de la base de données paléoécologiques.*

Summary

In this paper, we propose an approach for extracting (closed) frequent gradual patterns when the ordering of supporting objects matches the temporal order. This approach allows to reduce the quantity of mined patterns when the objects follow an temporal order relation. Experimental results obtained on the paleoecological data show the efficiency of our approach and the interpretation of results bring new knowledge to paleoecological experts.