Fouille de Motifs Graduels Fermés Fréquents Sous Contrainte de la Temporalité

Jerry Lonlac*,** Benjamin Negrevergne** Yannick Miras*** Aude Beauger***
Engelbert Mephu Nguifo*

*CNRS, UMR 6158, LIMOS, Université Clermont Auvergne, F-63173 Aubière, France
**LAMSADE, CNRS UMR 7243, Université Paris Dauphine
{benjamin.negrevergne}@dauphine.fr
***CNRS, UMR 6042, GEOLAB, Université Clermont Auvergne,
F-63000 Clermont-Ferrand
{jerry.lonlac_konlac, engelbert.mephu_nguifo, yannick.miras, aude.beauger}@uca.fr

1 Introduction

La fouille de motifs graduels a pour but la découverte de co-variations fréquentes de la forme "plus/moins X, plus/moins Y" 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.

2 Fouille de motifs graduels sous contrainte temporelle

Soit une base de données numériques Δ contenant un ensemble d'objets $\mathcal{D} = \{ d_1, ..., d_n\}$ décrit par un ensemble d'attributs $\mathcal{I} = \{ i_1, ..., i_m\}$. Nous dénotons par $d_j[i_k]$ la valeur de l'attribut i_k sur l'objet d_j. Un item graduel est défini sous la forme i^*, où i est un attribut de $\mathcal{I}$ et $* \in \{\geq, \leq\}$ est un opérateur de comparaison. Un itemset (motif) graduel $s = (i_1^{*_1}, ..., i_k^{*_k})$ est un ensemble non vide d'items graduels. Une séquence d'objets $\langle d_1, ..., d_s\rangle$ respecte s si $\forall p \in [1, s-1]$, $\forall l \in [1, k]$, nous avons $d_p[i_l] *_l d_{p+1}[i_l]$.

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. Dans ce travail, nous considérons la sémantique de variation qui définit le support d'un motif graduel s comme la longueur de la plus longue séquence d'objets respectant le motif s.

Soient s un motif graduel fréquent extrait de Δ, $L_s = \langle d_{l_1}, ..., d_{l_j} \rangle$ la plus longue séquence d'objets respectant s. Le motif s respecte l'ordre temporel des objets de Δ si on a l'inégalité suivante : $d_{l_1} < d_{l_2} < ... < d_{l_j}$.

Pour prendre en compte la contrainte de temporalité au cours du processus de fouille, nous intégrons une contrainte temporelle sur les variations des attributs dans l'encodage proposé dans Négrevergne et al. (2014). Cet encodage est défini comme suit : Soit $\mathcal{A} = \{ i_1^{\geq}, i_1^{\leq}, ..., i_m^{\geq}, i_m^{\geq} \}$ l'ensemble des variations d'attributs de $\mathcal{I}$. Les transactions dans la nouvelle base de données sont des paires d'objets $(d_j, d_{j'})$, $d_j, d_{j'} \in \mathcal{D}$, avec $j, j' \in [1, n]$ et $j < j'$. Nous dénotons par $t_{(d_j, d_{j'})}$ la transaction qui contient les variations pour tout attribut dans $\mathcal{A}$ entre les objets d_j et $d_{j'}$: pour tout attribut $i \in \mathcal{I}$, $i^{\geq} \in t_{(d_j, d_{j'})} \Leftrightarrow d_j[i] \leq d_{j'}[i]$, $i^{\leq} \in t_{(d_j, d_{j'})}$ sinon. La condition $j < j'$ permets d'imposer la contrainte de temporalité sur les variations d'attributs et constitue une petite optimisation comparée à l'encodage original.

3 Résultats expérimentaux

La figure 1 montre les résultats des tests effectués sur les bases de données paléoécologiques décrites dans Lonlac et al. (2017) contenant des attributs liés à l'eutrophisation. Ces résultats montrent que notre approche réduit le nombre de motifs générés par rapport à l'approche originale implémentée dans l'algorithme `Paraminer` (Négrevergne et al., 2014).

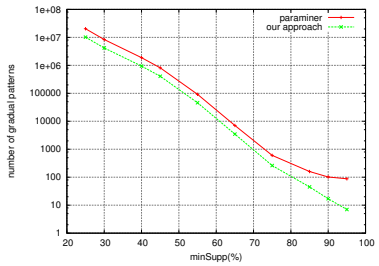

FIG. 1 – *Évaluation comparative de notre approche, avec variation du support minSupp.*

Interprétation des résultats : Les motifs extraits des données paléoécologiques sont pertinents car indicateurs de l'enrichissement trophique. Ils incluent des coévolutions d'indicateurs cohérents dans la mesure où les modèles d'indicateurs traduisent des coévolutions indiquant un statut trophique élevé des eaux du lac considéré. Ils permettent également de renforcer, voire de préciser le potentiel paléoécologique de certains indicateurs paléoécologiques.

Références

Lonlac, J., Y. Miras, A. Beauger, M. Pailloux, J.-L. Peiry, et E. Mephu (2017). Une approche d'extraction de motifs graduels (fermés) fréquents sous contrainte de la temporalité. In *EGC, 23-27 Janvier, Grenoble, France*, pp. 213–224.

Négrevergne, B., A. Termier, M. Rousset, et J. Méhaut (2014). Paraminer : a generic pattern mining algorithm for multi-core architectures. *Data Min. Knowl. Discov. 28*(3), 593–633.