

---

# Extraction de Motifs Graduels (Fermés) Fréquents Sous Contrainte de la Temporalité.

---

**Jerry Lonlac**<sup>1,2</sup>, Yannick Miras<sup>2</sup>, Aude Berger<sup>2</sup>, Marie Pailloux<sup>1</sup>,  
Jean-Luc Peiry<sup>2</sup>, Engelbert Mephu Nguifo<sup>1</sup>

<sup>1</sup>LIMOS - CNRS, Université Clermont Auvergne

<sup>2</sup>GEOLAB - CNRS, Université Clermont Auvergne

EGC'17 - Grenoble



# Plan

- 1 Motifs graduels
- 2 Extraction de motifs graduels : Etat de l'art
- 3 Notre approche : Gradualité + Temporalité
- 4 Cas d'étude et Expérimentation
- 5 Conclusion et Perspectives

# Sémantique d'un motif graduel

## Gradualité

- variation des valeurs d'éléments entre ensembles
- "plus X est A, plus Y est B"
- initialement utilisée dans le domaine du flou
  - systèmes experts
  - connaissances apportées

## Exemple

- plus l'âge d'un individu augmente, plus le montant de son salaire augmente

Nombreuses applications : Médecine, Biologie

# Sémantique d'un motif graduel

## Base d'exemple : indicateurs polliniques et leur abondance

Objet	Poacea (P)	Secale (S)	Rumex (R)
$t_1$	0	0	5
$t_2$	31	7	3
$t_3$	62	8	9
$t_4$	18	1	0
$t_5$	13	1	4
$t_6$	17	2	1
$t_7$	36	3	6

- voir le site web : <http://mobipaleo.univ-bpclermont.fr/>

# Sémantique d'un motif graduel

## Base d'exemple : indicateurs polliniques et leur abondance

Objet	Poacea (P)	Secale (S)
$t_1$	0	0
$t_5$	1	1
$t_4$	18	1
$t_2$	31	7
$t_3$	62	8
$t_6$	17	2
$t_7$	36	3

- nous travaillons sur les **valeurs** des items
- les comparaisons se font **entre les objets**

# Item graduel, Motif graduel

## Définition : item graduel

item de la forme  $i^*$  avec  $* \in \{+, -\}$  et  $dom(i)$  muni d'une relation d'ordre total

- $* = +$  signifie la valeur de  $i$  augmente
- $* = -$  signifie la valeur de  $i$  diminue

## Qu'est-ce que la variation ?

- $+$  correspond à  $\geq$
- $-$  correspond à  $\leq$
- $t_1[i] \geq t_2[i]$ , on écrit  $i^+$
- $t_1[i] \leq t_2[i]$ , on écrit  $i^-$

# Item graduel, Motif graduel

## Définition : item graduel

item de la forme  $i^*$  avec  $*$   $\in \{+, -\}$  et  $dom(i)$  muni d'une relation d'ordre total

- $*$  = + signifie la valeur de  $i$  augmente
- $*$  = - signifie la valeur de  $i$  diminue

## Base d'exemple : $P^+$

Objet	Poacea (P)
$t_1$	0
$t_2$	31
$t_3$	62
$t_4$	18
$t_5$	13
$t_6$	17
$t_7$	36

Objet	Poacea (P)
$t_1$	0
$t_5$	13
$t_6$	17
$t_4$	18
$t_2$	31
$t_7$	36
$t_3$	62

# Item graduel, Motif graduel

## Définition : motif (itemset) graduel

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

## Base d'exemple : $P^+R^-$

Objet	Poacea (P)	Rumex (R)
$t_1$	0	5
$t_5$	13	4
$t_6$	17	1
$t_4$	18	0

## Définition : motif graduel complémentaire

- soit  $s = (i_1^{*1}, \dots, i_k^{*k})$  et  $c$  tel que  $c(\geq) = \leq$  et  $c(\leq) = \geq$
- $c(s)$  désigne le complémentaire de  $s$
- exemple :  $c(P^+S^+R^-) = P^-S^-R^+$

# Etat de l'art

- **Berzal et al.** : An alternative approach to discover gradual dependencies. IJUFKS 2007
  - **Di Jorio et al.** : Mining frequent gradual itemsets from large databases. IDA 2009
    - **Ayouni et al.** : Mining closed gradual patterns. ICAISC 2010
  - **Laurent et al.** : Exploiting rank correlations for extracting gradual dependencies. FQAS 2009
  - **Do et al.** : Pglcm : efficient parallel mining of closed frequent gradual itemsets. ICDM 2010
- 
- différentes sémantiques (dépend de l'application considérée)
  - différents modes de calcul du support

# Motivation

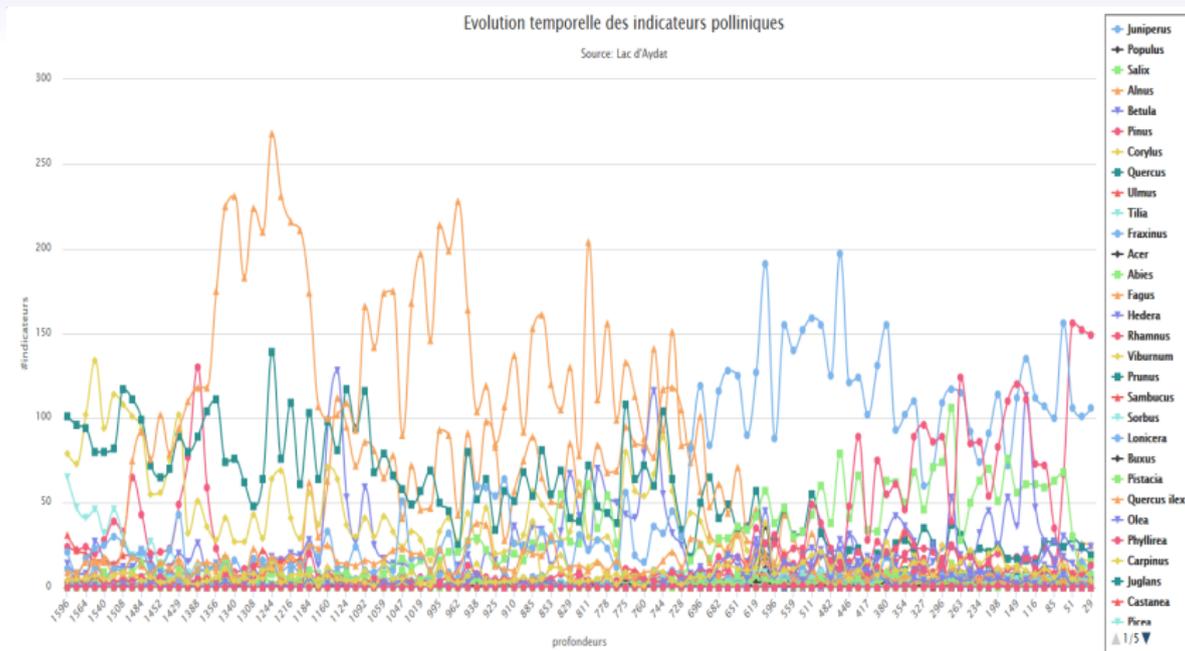
- certaines bases de données sont constituées d'objets munis d'une relation d'ordre temporel
- exemple : base de données Paléoécologiques

## Données Paléoécologiques (abondance pollinique)

Objet	Poacea (P)	Secale (S)	Rumex (R)
$t_1$	0	0	5
$t_2$	31	7	3
$t_3$	62	8	9
$t_4$	18	1	0
$t_5$	13	1	4
$t_6$	17	2	1
$t_7$	36	3	6

contrainte temporelle :  $t_1 < t_2 < \dots < t_7$

# Motivation : évolution de la biodiversité



décrire l'évolution de la biodiversité à partir des données Paléocéologiques

## Motivation : décrire l'évolution de la biodiversité

### Données paléoécologiques : $\Delta = \mathcal{T} \times \mathcal{I}$

- $\mathcal{T} = \{ t_1, \dots, t_n \}$  : ensemble de dates ordonnées
- $\mathcal{I} = \{ i_1, \dots, i_m \}$  : ensemble d'indicateurs paléoécologiques
- $\Delta(t, i)$  : nombre d'instances de l'indicateur  $i$  à la date  $t$

### Objectif

- extraire les groupements d'indicateurs paléoécologiques d'évolution

### Contraintes

- 1 temporalité
- 2 évolution entre périodes consécutives : haute résolution temporelle

motifs graduels sous contrainte de la temporalité

# Motifs graduels sous contrainte de la temporalité

## Notre approche

- considérer le formalisme proposé par [Berzal et al., 2007]
- prendre en compte au cours du processus de fouille
  - les contraintes 1 et 2
  - les cas des valeurs égales entre objets
- utiliser les approches de fouille d'itemsets classiques

## Avantages

- permet de fouiller dans les grands volumes de données
- permet d'extraire  $k$  – *itemsets* graduels ( $k > 3$ )

# Notre approche

## Définition 1 : $\Delta = \mathcal{D} \times \mathcal{I}$

- $\mathcal{I} = \{i_1, \dots, i_m\}$  : ensemble d'attributs numériques ou items
- $\mathcal{D} = \{d_1, \dots, d_n\}$  : ensemble d'objets
- $\Delta' = \mathcal{D}' \times \mathcal{I}'$ 
  - $\mathcal{D}' = \{d'_1, \dots, d'_{n-1}\}$ ,  $|\mathcal{D}'| = n - 1$
  - $\mathcal{I}' = \{i_1^>, i_1^<, i_1^=, \dots, i_m^>, i_m^<, i_m^=\}$ ,  $|\mathcal{I}'| = 3 \times |\mathcal{I}|$ ,
  - $\forall d'_k \in \mathcal{D}', i_j^{*p} \in d'_k \Leftrightarrow d_{k+1}[i_j] *p d_k[i_j]$ ,
- dans la suite : +, -, o correspondront respectivement à >, <, =

## Algorithme

- 1 transformer la matrice de données numériques  $\Delta$  en une matrice de données catégorielles  $\Delta'$  (en utilisant la définition 1)
- 2 rechercher tous les itemsets fermés fréquents de  $\Delta'$   
→ les itemsets (motifs) graduels (fermés) fréquents de  $\Delta$
- 3 rechercher les motifs graduels fermés ne contenant pas les items  $i^o$

## Notre approche

### Support d'un motif graduel [sous contrainte temporelle]

- $s = (i_1^{*1}, \dots, i_k^{*k})$  : motif graduel extrait sous contrainte temporelle
- $\mathcal{L}_s = \{L_1, \dots, L_s\}$  : séquences d'objets consécutifs respectant  $s$ .
  - $\forall i, j, i \neq j, L_i \cap L_j = \emptyset$
  - $Support(s) = \frac{\sum_{i=1}^s |L_i|}{|\mathcal{D}|}$ ,  $\mathcal{D}$  est la matrice de données initiales.
- $s$  est fréquent : support supérieur à un seuil minimal

### Base d'exemple : $P^+S^+$

Objet	Poacea (P)	Secale (S)
$t_1$	0	0
$t_2$	31	7
$t_3$	62	8
$t_4$	18	1
$t_5$	13	1
$t_6$	17	2
$t_7$	36	3

- $\mathcal{L}_{(P^+S^+)} = \{ \langle t_1, t_2, t_3 \rangle, \langle t_5, t_6, t_7 \rangle \}$
- $Support(P^+S^+) = \frac{6}{7}$
- $P^+S^+$  : sur 2 périodes consécutives,  $P$  croît et  $S$  croît.

# Application : la Paléoécologie

## Cas d'étude : Lac d'Aydat (Auvergne Rhône-Alpes)



- site touristique menacé d'eutrophisation
- reconstruire ces trajectoires écologiques sur 7 millénaires
- mieux connaître son état écologique actuel

## Données réelles utilisées

- trois jeux de données paléoécologiques ( $\Delta_1, \Delta_2, \Delta_3$ )
  - $\Delta_1$  : 111 objets, 87 variables (indicateurs d'anthropisation)
  - $\Delta_2$  : 57 objets, 178 variables (indicateurs statut trophique de l'eau)
  - $\Delta_3$  : 57 objets, 265 variables (anthropisation et eutrophisation)

construire les groupements d'indicateurs paléoécologiques d'évolution du Lac

# Application : la paléoécologie

## Expérimentations

- seuil de support minimal : 10%
- utiliser une version modifiée de *Apriori* [Agrawal et al., 1994]
- extraire les motifs graduels fermés fréquents à partir de  $\Delta_1, \Delta_2, \Delta_3$

## Post-traitement de l'ensemble des motifs extraits

- 1 unicité de la coévolution d'un motif fréquent  $s = \{i_1^{*1}, \dots, i_k^{*k}\}$ 
  - $\nexists s' = \{i_1^{*1'}, \dots, i_k^{*k'}\}$  tel que  $s \neq s'$  et  $s \neq c(s')$ .
- 2 au moins un indicateur direct d'activités humaines (IPA)
- 3 au moins un IPA + au moins un autre indicateur pollinique secondaire d'activités anthropiques (IPAS)

## Motifs graduels intéressants : impact anthropique

### Quelques motifs

- 1 Poaceae=+,Rumex.acetosella.type=+,Coprofilous.Fungi=+
- 2 Secale.type=+,Plantago.lanceolata=+,Ranunculus.acris.t=+
- 3 Secale.type=-,Plantago.lanceolata=-,Ranunculus.acris.t=-

### Nouvelles connaissances

- 1 discrimination des indicateurs secondaires d'anthropisation : association avec l'agriculture - association avec le pastoralisme
- 2 spécialisation des IPAS, coévolution préférentielle de certains IPAS avec des indicateurs d'agriculture ou du pastoralisme
- 3 nouveaux IPAS, coévolution préférentielle de certains IPAS avec des indicateurs encore peu utilisés : Polygonum, Apiaceae, Rosaceae

## Motifs graduels intéressants : eutrophisation de l'eau

### Cohérence des motifs extraits

- 1 confirmation des motifs par la littérature (notamment par un travail sur les lacs auvergnats [Rioual, 2000])
- 2 100% des motifs contenant des indicateurs d'eutrophisation

### Nouvelles connaissances

- 1 apport d'informations sur les preferenda écologiques
- 2 mise en évidence de motifs d'eutrophisation
  - espèces eutrophes, hypereutrophes vivant dans des eaux riches en matière organique
  - espèces à tendance mésotrophe à eutrophe
- 3 précision du signal paléocéologique : identification des motifs traduisant les conditions eutrophes voire hyper-eutrophe

# Conclusion et Perspectives

## Conclusion

- recherche de motifs graduels dans des **données munies d'une relation d'ordre temporel**
- utilisation des algorithmes existants d'extraction d'itemsets
- expérimentations sur des jeux de données réelles
- contribution à la résolution d'un problème scientifique

## Perspectives

- mesure de qualité des motifs extraits (taux de variations)
- étude des différents gaps de variations (entre objets éloignés)
- recherche des motifs extraits sur d'autres sites d'étude (fouille multi-vues)

# Merci de votre attention

## Questions ?

- voir le site web : <http://mobipaleo.univ-bpclermont.fr/>
- remerciements :

