Fouille de données environnementales

Florence Le Ber

florence.leber@engees.unistra.fr
http://engees.unistra.fr/~fleber/

LHYGES–ENGEES & LORIA

Séminaire LIRIS – Lyon, 30 septembre 2010

Plan

Généralités

- L'objectif du processus d'**extraction de connaissances dans les bases de données (ECBD ou ECD)** est d'extraire dans des grands volumes de données des **unités de connaissances significatives** et **réutilisables**.
- Les **données** sont disponibles en quantité importante et en qualité variable, sans utilisation particulière et précise a priori.
- Les unités de connaissances extraites peuvent avoir plusieurs formes, par exemple : **classes d'individus**, **motifs** (ensemble de propriétés) **règles d'association**, **dépendances fonctionnelles**.

Le processus d'ECBD

Données
↓ Sélection et préparation des données
↓ Transformations : nettoyage et mise au format
Données préparées
↓ Opérations de fouille de données
↓ Méthodes numériques et/ou symboliques
Unités extraites
↓ Interprétation / Évaluation
↓ Représentation des unités extraites
Unités de connaissances
↓
(Système à base de connaissances)

Le processus d'ECBD est itératif et interactif

L'analyste dans le processus d'ECBD

Le processus d'ECBD est **itératif** et **interactif** : il placé sous le contrôle d'un expert du domaine appelé l'**analyste**.
Cet expert oriente le processus de fouille selon ses **propres connaissances** et selon les connaissances du domaine voire une **représentation du modèle du domaine** (**ontologie**).
Il interprète les éléments extraits pour faire émerger des unités de connaissances réutilisables et relance le processus avec de nouveaux paramètres.

Méthodes en fouille de données : Méthodes symboliques

- La classification par treillis.
- La recherche de motifs fréquents et l'extraction de règles d'association.
- La classification par arbres de décision.
- Les méthodes inductives en apprentissage : à partir d'instances, à partir d'exemples, ou à partir de cas.
- Les méthodes de recherche d'information et d'interrogation de bases de données.

Méthodes en fouille de données : Méthodes numériques

- Les méthodes statistiques et d'analyse des données.
- Les modèles de Markov cachés, conçus et mis au point à l'origine pour la reconnaissance de formes (parole, image, caractères).
- Les réseaux bayésiens pour la recherche de causalités.
- Les réseaux de neurones.
- Les algorithmes génétiques.

Treillis

Définition

Un treillis $(\mathtt{E}, \leq)$ *est un ensemble ordonné tel que chaque couple d'éléments* $(\mathtt{x}, \mathtt{y})$ *possède un* ***supremum*** *noté* $\mathtt{x} \vee \mathtt{y}$ *et un* ***infimum*** *noté* $\mathtt{x} \wedge \mathtt{y}$.

- L'ensemble $2^{\mathtt{E}}$ des parties d'un ensemble $\mathtt{E}$ muni de la relation d'inclusion est un treillis.
- L'ensemble **N** des entiers naturels muni de la relation de divisibilité est un treillis : $\mathtt{x} \vee \mathtt{y} = \mathtt{pgcd}(\mathtt{x}, \mathtt{y})$ et $\mathtt{x} \wedge \mathtt{y} = \mathtt{ppmc}(\mathtt{x}, \mathtt{y})$.

La connexion de Galois

Définition

*La **connexion de Galois** $(\mathtt{f}, \mathtt{g})$ d'un contexte $\mathtt{O} \times \mathtt{A}$ se définit comme suit :*

- *La fonction $\mathtt{f} : 2^{\mathtt{O}} \longrightarrow 2^{\mathtt{A}}$ associe à un sous-ensemble d'objets $\mathtt{X}$ de $\mathtt{O}$ le sous-ensemble d'attributs (items ou propriétés) communs qu'ils partagent $\mathtt{Y}$ de $\mathtt{A}$ (**intension** de $\mathtt{X}$).*
- *De façon duale :*
  *La fonction $\mathtt{g} : 2^{\mathtt{A}} \longrightarrow 2^{\mathtt{O}}$ associe à un sous-ensemble d'attributs $\mathtt{Y}$ de $\mathtt{A}$ le sous-ensemble d'objets $\mathtt{X}$ de $\mathtt{A}$ qui possèdent ces attributs (**extension** de $\mathtt{Y}$).*

Exemple

Soit le contexte binaire (O, T, I) : O un ensemble d'objets, T un ensemble d'attributs et I une relation associant à tout objet de O les attributs qu'il possède.

Objets / Attributs	a	b	c	d	e
o1		x	x		x
o2	x		x	x	
o3	x	x	x	x	
o4	x			x	
o5	x	x	x	x	
o6	x		x	x	

$f(\{o_1\}) = \{b, c, e\}$ et $g(\{b, c, e\}) = \{o_1\}$
$f(\{o_1, o_2\}) = \{c\}$ et $g(\{c\}) = \{o_1, o_2, o_3, o_5, o_6\}$
$g(\{a, c\}) = \{o_2, o_3, o_5, o_6\}$ et $f(\{o_2, o_3, o_5, o_6\}) = \{a, c, d\}$

Treillis de Galois

Les fonctions $\mathtt{h} = \mathtt{g}(\mathtt{f})$ et $\mathtt{h}' = \mathtt{f}(\mathtt{g})$ sont des opérateurs de **fermeture** : ils sont croissants, extensifs, et idempotents.

- Un **fermé** $\mathtt{X}$ de $\mathtt{h}$ vérifie $\mathtt{h}(\mathtt{X}) = \mathtt{X}$; un **fermé** $\mathtt{Y}$ de $\mathtt{h}'$ vérifie $\mathtt{h}'(\mathtt{Y}) = \mathtt{Y}$.
- $\mathtt{h}$ et $\mathtt{h}'$ permettent de construire le **treillis de Galois** du contexte $\mathtt{O} \times \mathtt{A}$.

Définition

Le treillis de Galois est défini comme le produit des deux treillis isomorphes $\mathtt{L}_\mathtt{O} \times \mathtt{L}_\mathtt{A}$, *où* $\mathtt{L}_\mathtt{O}$ *est le treillis des fermés pour* $\mathtt{h}$ *— ou* ***treillis des extensions*** *— et* $\mathtt{L}_\mathtt{A}$ *le treillis des fermés pour* $\mathtt{h}'$, *ou* ***treillis des intensions***.

Concepts

Dans un treillis de Galois ou **treillis de concepts**, les concepts sont des couples $\mathtt{C}_k = (\mathtt{E}_k, \mathtt{I}_k)$, définis par une une **extension** $\mathtt{E}_k$ et une **intension** $\mathtt{I}_k$:

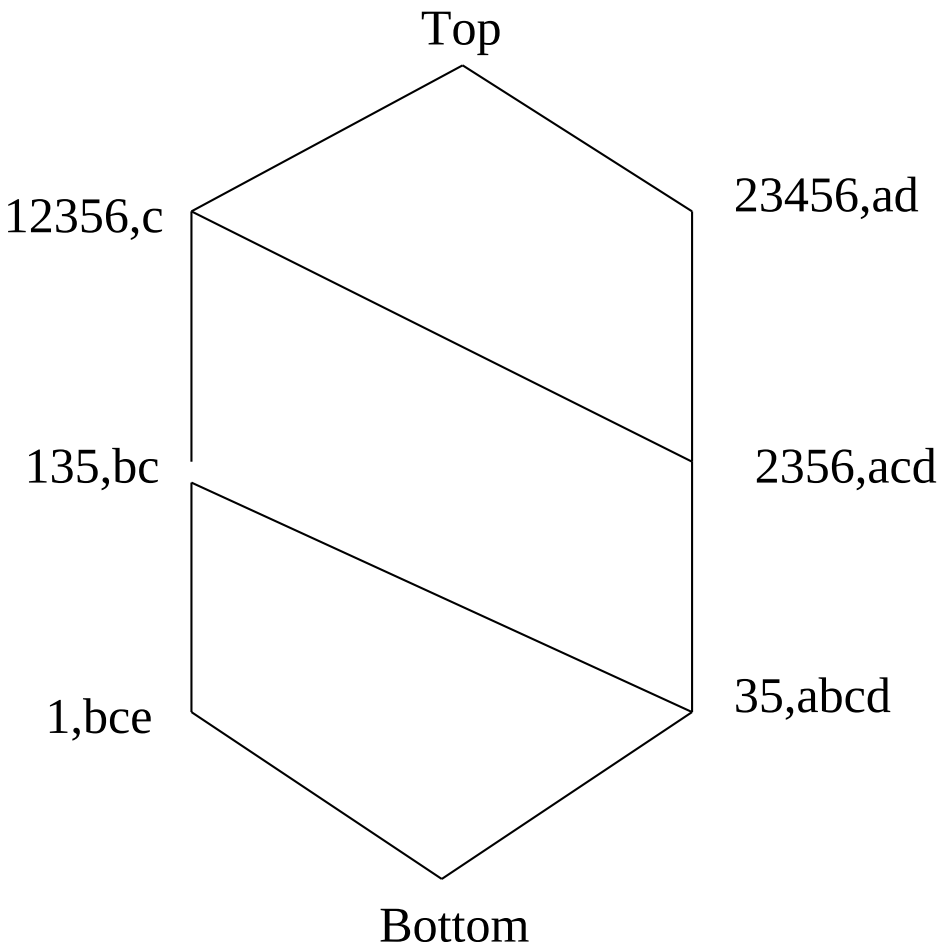
- L'**extension** $\mathtt{E}_k$ représente l'ensemble des objets recouverts par le concept, ou objets qui possèdent les attributs de $\mathtt{I}_k$.
- L'**intension** $\mathtt{I}_k$ représente l'ensemble dual des attributs du concept, ou les attributs possédés par les objets de $\mathtt{E}_k$.

Définition

L'ordre partiel dans un treillis de concepts, noté $\sqsubseteq$, est défini par :

$(\mathtt{E}_1, \mathtt{I}_1) \sqsubseteq (\mathtt{E}_2, \mathtt{I}_2)$ *ssi* $\mathtt{E}_1 \subseteq \mathtt{E}_2$, *ou de façon duale,* $\mathtt{I}_2 \subseteq \mathtt{I}_1$.

Exemple (suite)



Types d'attributs

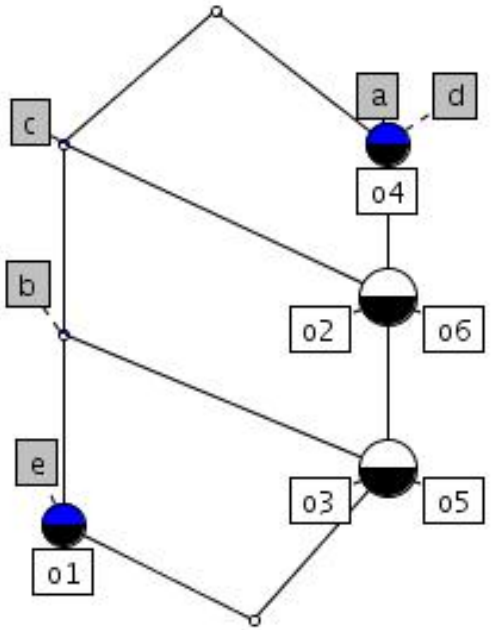
Attribut (ou propriété) propre

un attribut est **propre** à un concept s'il n'est défini dans aucun des ascendants du concept ("concept le plus haut où apparaît l'attribut").

Attribut (ou propriété) hérité

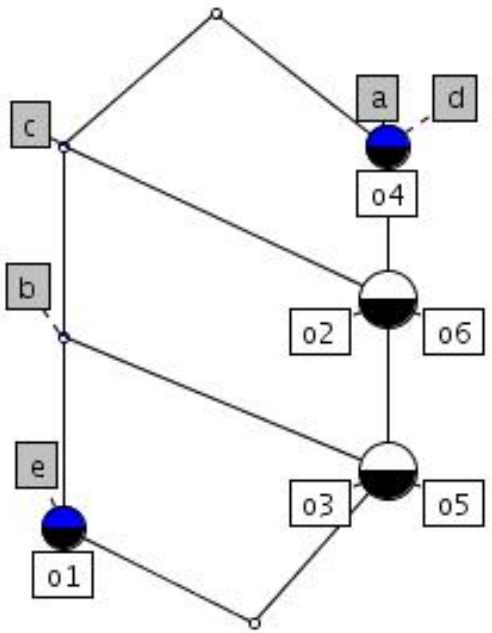
un attribut est **hérité** par un concept s'il est défini dans un des ascendants du concept ("concept inférieur au concept de définition pour l'ordre du treillis").

Exemple



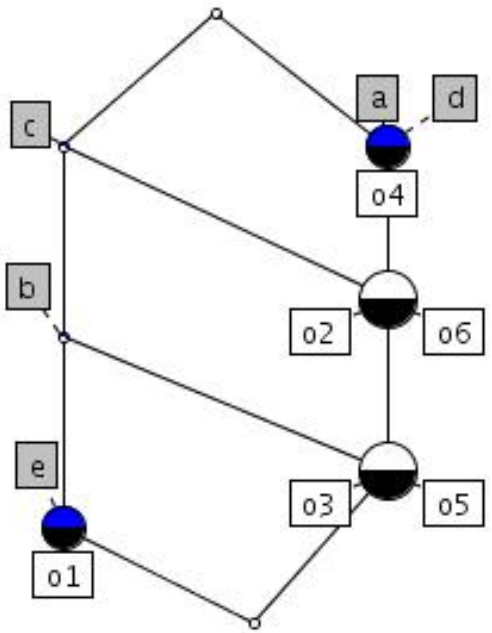
- `c` est un attribut propre du concept (`12356`, `c`), `a` et `d` sont des attributs propres du concept (`23456`, `ad`), `b` est un attribut propre du concept (`135`, `bc`).
- `c` est un attribut hérité par (`135`, `bc`), `a`, `c` et `d` sont des attributs hérités par (`2356`, `acd`), ...

Règles extraites du treillis : les implications mutuelles entre attributs propres



- Deux attributs propres sont **équivalents**, par exemple $\mathtt{a} \longleftrightarrow \mathtt{d}$ pour $(\mathtt{23456}, \mathtt{ad})$.

Règles extraites du treillis : les attributs propres impliquent les attributs hérités



- Un attribut propre **implique** un attribut hérité, par exemple $\mathtt{b} \longrightarrow \mathtt{c}$ pour $(\mathtt{135}, \mathtt{bc})$ et $\mathtt{e} \longrightarrow \mathtt{bc}$ pour $(\mathtt{1}, \mathtt{bce})$.

Contextes multi-valués

(Ganter et Wille, 1999)

$K := (O, T, M, I)$

- O : un ensemble d'objets
- T : un ensemble d'attributs ou propriétés
- M : un ensemble de modalités ou valeurs d'un attribut
- $I \subseteq O \times T \times M$ une relation telle que $(o,t,m) \in I$ et $(o,t,n) \in I$ implique $m = n$.
- La notation $(o,t,m) \in I$ (ou $t(o) = m$) signifie que l'attribut t prend la valeur m pour l'objet o.

Traitement par échelonnage conceptuel (*conceptual scaling*) : le contexte est ramené à un contexte binaire avec autant d'attributs binaires qu'il y a de modalités dans le contexte d'origine.

Contexte flou

(Bělohlávek, 1999)

- A : un ensemble de degrés de vérité.
- $K := (O, T, I)$: un contexte flou avec $I : O \times T \rightarrow A$ une relation floue entre O et T à valeur dans A.
- Un degré (ou affinité) $I(o, t) \in A$ représente le degré auquel l'objet o possède l'attribut t.

Ces contextes nécessitent la définition de connexions spécifiques, les connexions floues : à un ensemble d'objets on associe l'ensemble flou de leurs attributs communs chacun assorti du degré de vérité minimal pour cet ensemble d'objets. À un ensemble flou d'attributs m^{β}, on associe les objets qui possèdent chaque attribut avec un degré de vérité supérieur à β.

Contexte flou multi-valué

(Bertaux *et al.*, 2009)

$K := (O, T, M, A, I)$

- O : un ensemble d'objets
- T : un ensemble d'attributs ou propriétés
- M : un ensemble de modalités ou valeurs d'un attribut
- A : un ensemble d'affinités ou degré de vérité (sur les modalités des attributs)
- $I : O \times T \times M \rightarrow A$: une relation floue telle que : $I(o, t, m) = a$ signifie que l'objet o possède la modalité m de l'attribut t avec un degré de vérité a
- Un objet peut donc posséder plusieurs modalités d'un attribut. La somme de leurs degrés de vérité vaut 1.

Le projet Indices

Projet subventionné par l'Agence de l'Eau Rhin Meuse (2006-10, resp. C. Grac, LHyGeS–ENGEES), intitulé :

« Comparaison d'indices biologiques pour l'évaluation de la qualité des cours d'eau »

Objectifs :

- Mise en relation des 5 indices biologiques et des pressions sur un ensemble de stations de la plaine d'Alsace
- Études des traits de vie des taxons (thèse d'Aurélie Bertaux)
- Construction d'un système d'évaluation globale de l'état écologique des stations de rivières

Traits biologiques des macrophytes

Les traits

Connaissances générales sur les caractéristiques physiques et biologiques des plantes.

Le jeu de données

- 50 objets (espèces)
- 10 attributs (traits biologiques)
- 35 modalités
- 100 affinités

Trait "reproduction végétative" (7 espèces)

Traits	reproduction végétative			
Modalités	*bulbe ou tubercule*	*Rhizome ou stolon*	*bulbille, turion ou apex dormant*	*fragments non spécialisés*
ALIP	0	100	0	0
BERE	0	66	0	33
CALO	0	0	0	100
CERD	0	0	50	50
ELON	0	0	33	66
GROD	40	40	0	20
PTNA	0	50	25	25

Traitement sous forme d'histogrammes

Le tableau des traits est un tableau de données multivaluées (plusieurs modalités) floues (plusieurs affinités). Afin de le traiter par les treillis de Galois, il faut le transformer en tableau binaire et/ou définir une connexion de Galois adaptée.

Un objet est alors associé à un trait-histogramme, exemple : V-40-40-0-20 pour GROD (*Groenlandia densa*) ou V-0-50-25-25 pour PTNA (*Potamogeton natans*), autrement dit un vecteur de valeurs, dont la somme vaut 100, pour chaque trait.

Deux objets appartiennent à un même concept s'ils ont un même ensemble de traits-histogrammes, ce qui est rare (on a un tableau binaire creux). Pour avoir plus de résultats, on modifie la connexion de Galois en une connexion floue : deux objets appartiennent à un même concept si leurs traits-histogrammes ne sont pas trop "distants".

Connexion de Galois avec similarité

O est l'ensemble des objets (macrophytes), Θ l'ensemble des traits-histogrammes, $\theta(o)$ le trait-histogramme d'un objet o et s un seuil de similarité défini sur les traits-histogrammes,

$$f : X \subseteq O \rightarrow Y = \begin{cases} \{\theta \in \Theta | \min_{o \in X} \theta(o) \leq \theta \leq \max_{o \in X} \theta(o)\} \text{ si } |\max_{o \in X} \theta(o) - \min_{o \in X} \theta(o)| \leq s \\ \emptyset \text{ sinon} \end{cases}$$

$$g : Y \subseteq \Theta \rightarrow X = \begin{cases} \{o \in O | \min_{\theta \in Y} \theta \leq \theta(o) \leq \max_{\theta \in Y} \theta\} \text{ si } |\max_{\theta \in Y} \theta - \min_{\theta \in Y} \theta| \leq s \\ \emptyset \text{ sinon} \end{cases}$$

Différentes valeurs de seuils donnent différents ensembles de concepts (taille des extensions variables). La différence min max peut être évaluée sur l'ensemble des traits ou trait par trait, voire modalité par modalité.

Exemple

Pour le contexte réduit : ALIP (V-0-100-0-0), ELON (V-0-0-33-66), GROD (V-40-40-0-20), PTNA (V-0-50-25-25).
Sans prise en compte du seuil :

$$f(X = \{ALIP, ELON\}) = [\text{V-0-0-0-0}, \text{V-0-100-33-66}]$$
$$g(Y = [\text{V-0-0-0-0}, \text{V-0-100-33-66}]) = \{ALIP, ELON, PTNA\}$$

$$|\max_{o \in X} \theta(o) - \min_{o \in X} \theta(o)| = |\text{V-0-100-33-66} - \text{V-0-0-0-0}| = 200$$

- Si on considère un seuil propre pour chaque modalité, par exemple $s = 50$ alors la deuxième modalité ne le respecte pas et le concept est rejeté.
- Si on considère un seuil global sur le trait, proportionnel au nombre de modalités, par exemple ici $s = 50 \times 4$, alors le concept peut être conservé.

Résultats

Sélection de groupes de macrophytes possédant des distributions de modalités proches pour un ensemble de traits (avec M. Trémolières, LHyGeS–UdS).

Macrophytes	Reproduction végétative			
CALO CERD ELOC ELON MYRS MYRV	0-0	0-40	0-50	33-100
BERE MENA MYOP NASO PHAA VERB	0-0	33-100	0-0	0-66
ALIP IRIP MENA MYOP PHAA SAGS SEFC	0-25	25-100	0-50	0-50
NUPL PTCR PTLU PTNA PTNO RANC RANU	0-0	0-100	0-40	0-100
HOTP JUNA MYOP NASO VERB	0-0	33-100	0-0	0-66
CERD GROD OENF PTPE RANU	0-40	0-40	0-50	20-100
CALO CALP PTPE RANT VERA ZANP	0-33	0-66	0-0	33-100
MYOP NYMA PTCR PTNA VERB	0-0	40-100	0-40	0-50

Suite du travail : mise en relation avec les pressions *via* les stations de prélèvement (analyse relationnelle de concepts).
Même approche sur les autres groupes de taxons (invertébrés, etc.)

Étude des valeurs d'indices biologiques

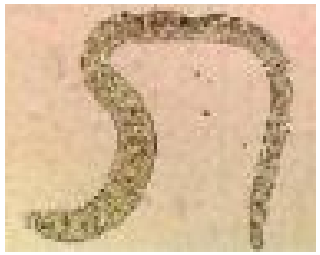
Objectifs

- Étudier et comparer les valeurs des différents indices biologiques sur un ensemble de stations subissant des pressions variées
- Être capable de choisir le meilleur indice pour évaluer une pression
- Proposer un outil d'évaluation globale de l'état écologique d'une station

IBGN : invertebrés

IBD : diatomées

IOBS : oligochètes



IBMR : macrophytes

IPR : poissons

Données

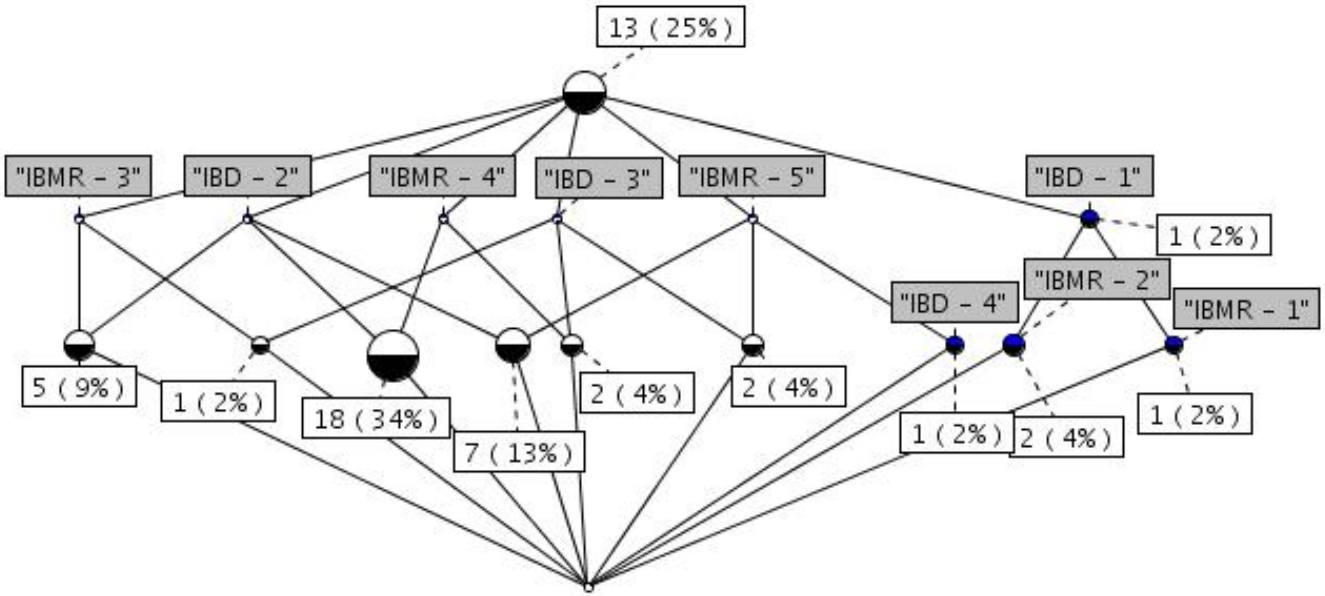
Collecte de données sur quarante stations de la plaine d'Alsace

Code station	IBGN	IBMR	IOBS	IBD	IPR
BREI001	3	5	incalculable	3	5
BRUN002	2	4	3	3	2
HORG001	3	3	5	3	3
STEI002	1	3	3	2	1
...	⋮	⋮	⋮	⋮	⋮

Chaque indice renseigne sur un ou des compartiments du milieu aquatique : par exemple les oligochètes vivent dans les sédiments. Un indice IOBS élevé indique donc un mauvais état des sédiments, alors que l'IPR ou l'IBD rendent compte de l'état de l'eau.

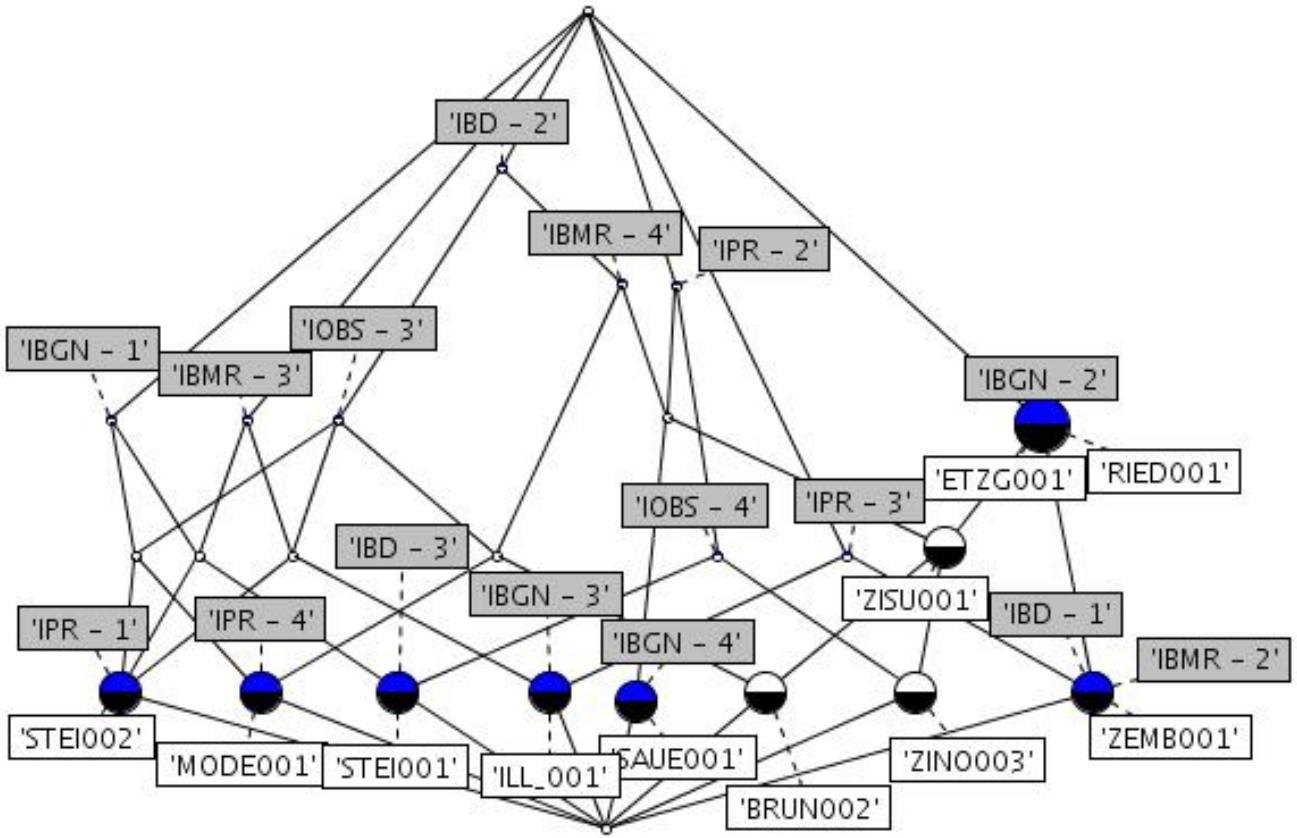
Construction de treillis par valeurs d'indices

Exemple : combinaison de valeurs pour les indices IBD et IBMR



Construction de treillis par degré d'altération

Classement des stations en fonction de leur degré d'altération (= moyenne des classes d'indices) et étude par les treillis : mise en évidence des classes d'indices communes ou non pour un même degré d'altération



Suite du travail : pour l'expert

Qualifier (« nommer ») les concepts

À un couple ({stations}, {classes de qualité}) est associée une qualification de l'état du milieu. Par exemple, les stations possédant les classes de qualité

(IBGN=[1,2], IOBS=3, IBMR=3, IBD=[4,5])

sont qualifiées par :

"début de dégradation des sédiments, forte dégradation physico-chimique au moins liée à un niveau trophique moyen, mais hors matière organique, bon potentiel de résilience général et possibilité de résilience sur les sédiments".

Suite du travail : sur les treillis

Construire un système de classement

Classer une nouvelle station revient à rechercher s'il existe dans le treillis un concept ayant les mêmes attributs que la station à classer. Si ce concept existe, la station peut y être rattachée, sinon, la structure hiérarchique du treillis permet d'obtenir une réponse « approchée », constituée par les concepts les plus proches (plus spécifiques ou plus généraux) du concept visé.

→ Utiliser un algorithme incrémental de construction de treillis (Carpineto et Romano, 2004). On n'est pas obligé de bâtir tout le treillis : on peut déterminer le concept-requête à partir du contexte (par fermeture) et ne construire – pour la visualisation – que la partie du treillis centrée autour de ce concept.

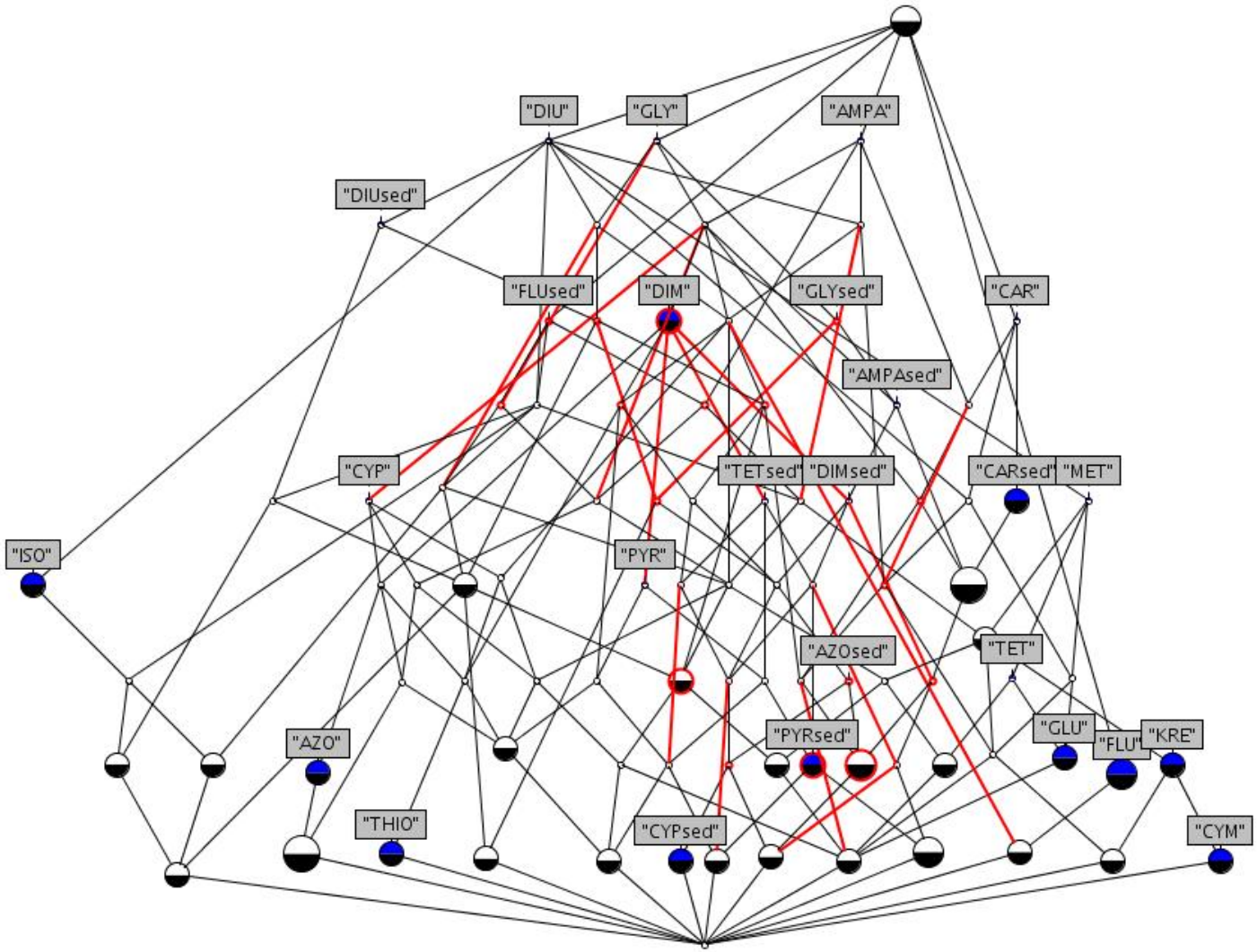
Étude des relations entre pesticides et invertébrés

Travail conjoint avec G. Imfeld (LHyGeS–CNRS) et A. Bertaux : suivi temporel et spatial des teneurs en pesticides et invertébrés dans le bassin d'orage de Rouffach (Haut Rhin) : relevés et analyses sur 4 zones, 10 dates. Tableaux numériques transformés au format binaire (présence/absence).

	DIU	THIO	ISO	GLY	AMPA	GLU	CYM	DIM	KRE	PYR
1_1	0	0	0	0	0	0	0	0	0	0
1_2	0	0	0	0	0	0	0	0	0	0
1_3	0	0	0	0	0	0	0	0	0	0
2_1	1	0	0	1	1	0	0	0	0	0
2_2	1	0	0	1	1	0	0	0	0	0

Construction de treillis pour les pesticides

Visualisation des groupes de sites (zone, date) partageant la présence de pesticides



Extraction d'implications sur les espèces

< 39 > { } → TUBtub CHIpsec
< 38 > TUBtub CHIchi CHIpsec → NAIchaeto
< 38 > NAIchaeto TUBtub CHIpsec → CHIchi
< 25 > TUBtub CHIparacla CHIpsec → NAIchaeto CHIchi
< 25 > TUBlimno TUBtub CHIpsec → NAIchaeto CHIchi
< 22 > TUBtub CHIpsec METRE → NAIchaeto CHIchi
< 18 > Nainais TUBtub CHIpsec → NAIchaeto CHIchi
< 15 > TUBtub CHIpsec NEMA → TUBner
< 15 > LUMlum TUBtub CHIpsec → TUBner
< 13 > TUBtub CHIpsec CHItanytar → NAIchaeto TUBner CHIchi NEMA
< 13 > TUBpelo TUBtub CHIpsec → NAIchaeto CHIchi
< 11 > TUBtub CHIpsec PLAgyr → NAIchaeto CHIchi
< 11 > NAIchaeto TUBner TUBtub CHIchi CHIpsec LYMgal NEMA → CHItanytar
< 10 > NAIchaeto Nainais TUBner TUBtub CHIchi CHIpsec → LUMlum CHItanytar NEMA

Suite du travail

- Transformation des tableaux en classes ou valeurs d'intervalles.
- Treillis de Galois sur les intervalles (connexion spécifique). Cf. Travaux de thèse de Z. Assaghir et M. Kaytoue (en cours) dans l'équipe Orpailleur au LORIA.
- Comparaison avec les approches statistiques

Bibliographie


R. Bělohlávek, "Fuzzy Galois Connections", Math. Logic Quaterly, vol. 45, 1999, pp. 497-504.


A. Bertaux. Treillis de Galois pour les contextes multi-valués flous. Application à l'étude des traits de vie en hydrobiologie. Thèse de doctorat de l'Université de Strasbourg, octobre 2010.


A. Braud, C. Grac, S. Pristavu, E. Dor et F. Le Ber. Une démarche fondée sur les treillis de Galois pour l'aide à la qualification de l'état des milieux aquatiques. In : *Actes du 2ème Atelier "Systèmes d'Information et de Décision pour l'Environnement" - SIDE 2009, Toulouse*, pp. 94-105, mai 2009.


A. Bertaux, F. Le Ber, A. Braud et M. Trémolières. Identifying ecological traits : a concrete FCA-based approach. In : *7th International Conference on Formal Concept Analysis, ICFCA 2009, Darmstadt*, Springer, LNAI 5548, pp. 224-236, 2009.


B. Ganter et R. Wille. Formal Concept Analysis : Mathematical foundations, Springer, 1999.


A. Napoli, A smooth introduction to symbolic methods in knowledge discovery. In : *Categorization in Cognitive Science*, H. Cohen et C. Lefebvre (eds), Elsevier, 2006.