

Identification de traits écologiques : une démarche concrète basée sur l'Analyse de Concepts Formels

Aurélie Bertaux, Florence Le Ber, Agnès Braud, Michèle Trémolières

Ecole Nationale du Génie de l'Eau et de l'Environnement de Strasbourg
Laboratoire d'Hydrologie et de Géochimie de Strasbourg UMR 7517

Laboratoire des Sciences de l'Image, de l'Informatique et de la Télédétection UMR 7005
Fouille de Données, Bioinformatique Théorique

GDR-I3 - Strasbourg - 2010



Laboratoire
d'hydrologie et
de géochimie
de Strasbourg
UMR 7517

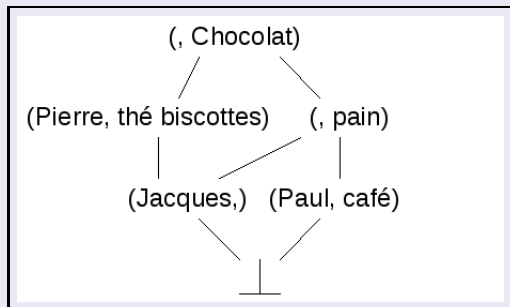


Treillis de Galois

Contexte binaire

- Données binaires d'*attributs* possédés par des *objets* et inversement.
- $f(\text{Pierre}) = \text{thé chocolat biscottes}$ et $g(\text{café}) = \text{Paul}$
- Le couple (f,g) : *correspondance de Galois* permettant d'obtenir les *concepts*.

Treillis de Galois



Exemple : habitudes de petit déjeuner

	café	thé	chocolat	pain	biscottes
Pierre	0	1	1	0	1
Paul	1	0	1	1	0
Jacques	0	1	1	1	1

Lattices for complex data

Complex data are often dealt as :

- Many-valued contexts using scaling methods (complexity on the attributes)
- Fuzzy contexts for (on the relation)

Actually the data we deal with are more complex. So we define :

- Fuzzy many-valued contexts (complexity on attributes or relation)
- Histogram scaling

We apply these definitions to hydrobiological problem using FCA-based approach.

1 Formal Concept Analysis

- Classical approaches
 - Many-valued context
 - Fuzzy context
- Complex data
- Definitions
 - Fuzzy many-valued context
 - Histogram scaling

2 Hydrobiological application

- Concept analysis
 - Concept selection
 - Expert interpretation : biological to ecological traits
- Validation
 - Dataset upgrading
 - Validation of the method
 - Applicative validation

3 Conclusion and future work

- Conclusion
- Future work

Many-valued context (complexity on the attributes)

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

- O : set of Objects
- T : set of attributes called *Traits*
- M : set of Modalities : values of a trait
- $I \subseteq O \times T \times M$
- The notation $(o, t, m) \in I$ (or $t(o) = m$) means that the t attribute has the m modality for the o object.

Fuzzy context (complexity on the relation)

- A : set of truth degrees.
- $K := (O, T, I)$: fuzzy context with $I : O \times T \rightarrow A$ a fuzzy relation between O and T .
- A degree $I(o, t) \in A$ is a degree to which the o object has the t trait.

A formal context is a specific case of a fuzzy context where $I : O \times T \rightarrow \{0, 1\}$.

Dataset

Complex domain

General knowledge about biological characteristics of macrophytes in the Alsace plain, collected in the literature.

Particularities

- 50 objects (species)
- 10 attributes (biological traits)
- 35 modalities
- Affinities (%)

Vegetative reproduction trait for a 7 species subset

Traits	Vegetative reproduction			
Modalities	<i>bulb or tubercle</i>	<i>Rhizome or stolon</i>	<i>bulbil, turion or dormant apex</i>	<i>non specialized fragments</i>
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

Fuzzy many-valued context (complexity on attributes and relation)

Definition

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

- A : set of the Affinities
- The notion of *fuzzy many-valued context* extends many-valued context definition
- m and n of t belong to o with different degrees

Traits	Vegetative reproduction			
Modalities	bulb or tubercle	Rhizome or stolon	bulbil, turion or dormant apex	non specialized fragments
GROD	40	40	0	20

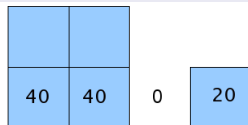
Histogram Scaling

What we need

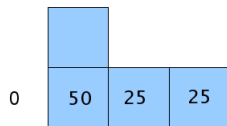
We need to obtain a binary context, so we have to transform the attributes. The goal is :

- to obtain attributes usable by classical binary lattices algorithms
- keep the information of repartition of the species between modalities

Example of Histograms



GROD



PTNA

bulb or Rhizome bulbil, turion non
tubercle or stolon or dormant specialized
apex fragments

fbox

Histogram Scaling

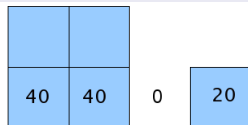
Definition

In order to deal with $K := (O, T, M, A, I)$ context,

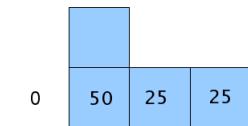
- definition of a H set of histograms attributes
- $n_M(t)$: number of modalities for the t trait. Example $n_M(\text{Vegetative reproduction}) = 4$
- An $h_t \in H$ histogram is described as :
 - a letter to qualify the considered t trait. Example V for Vegetative reproduction
 - the $n_M(t)$ affinities of the considered object for the $n_M(t)$ modalities of t

Example : V-40-40-0-20 or V-0-50-25-25

Example of Histograms



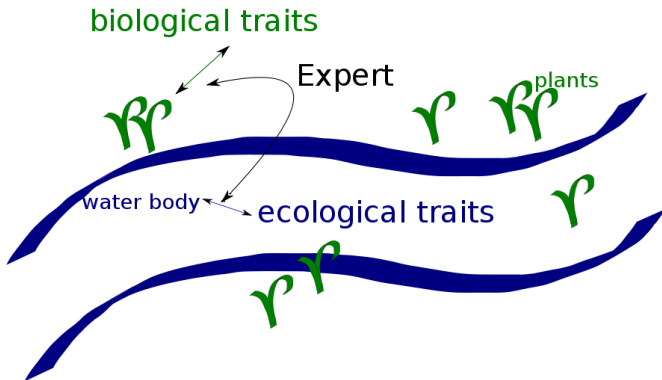
GROD



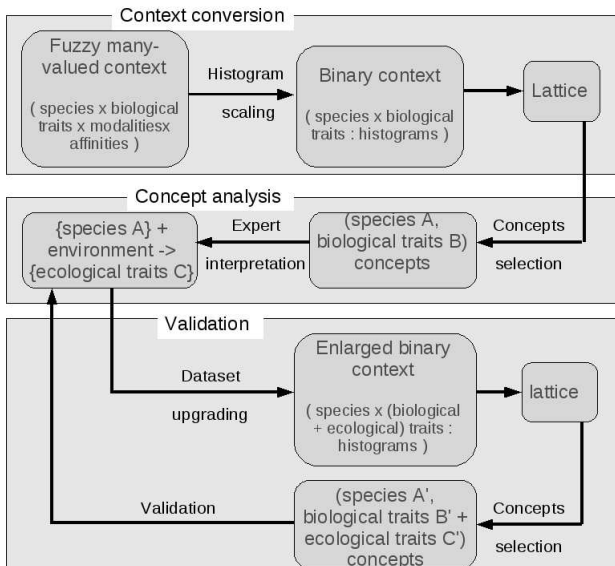
PTNA

bulb or Rhizome bulbil, turion non
tubercle or stolon or dormant specialized
apex fragments

What we do



How we do : the whole process



Concept selection

Meaningful concepts for biologists : between 3 and 5 attributes, i.e. between 3 and 7 species. Analysis of two concepts :

- Concept 1 : (JUNA SEFC TYPL, L100-0 H0-100 P100-0-0 V0-100-0-0 D100-0-0)
 - annual flowering,
 - a phenology during the vegetative period only,
 - Perennial organs (aerial or underground),
 - a Vegetative reproduction by rhizomes or by stolons ,
 - high Dispersion (with small flying seeds)
- Concept 2 : (CERD PHAA PTCR PTLU PTNA UTRV, H0-100 F0-0-0-100 P100-0-0)
 - a phenology during the vegetative period only,
 - an high Flexibility ($> 300^\circ$),
 - Perennial organs (aerial or underground)

Expert interpretation : biological to ecological traits

Expert knowledge

Select the most appropriate ecological traits among the many existing ones.
Association of the selected extents and information about environment where species live.

- Concept 1 : (JUNA SEFC TYPL, L100-0 H0-100 P100-0-0 V0-100-0-0 D100-0-0) : live in mesotrophic to eutrophic water
- Concept 2 : (CERD PHAA PTCR PTLU PTNA UTRV, H0-100 F0-0-0-100 P100-0-0) : live in calm water surfaces

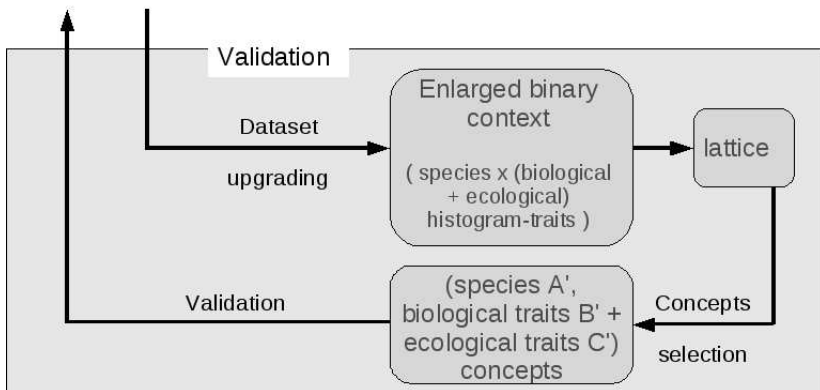
Ecological traits and there modalities selected

- water level stability : stable, fluctuations, occasionally
- resistance to flow : no tolerance, weak, medium, strong
- tolerance to organic matter : <10%, 10-40%, >40%
- tolerance to sedimentation : suffocated plant, medium root, strong root
- trophic status : oligotrophic, mesotrophic, eutrophic, hypertrophic

Validation step

Questions to answer

- Does the addition of the ecological traits in the context keep the context stable i.e. does it preserve previous concepts ?
- Do the ecological traits in the intents correspond to expert analysis ?



Dataset upgrading

Upgrading process

- Every specie is concerned by these ecological traits, not only the ones belonging to the concept(s) which pointed out the ecological trait.
- Ecological traits and their modalities added to the dataset
- 46 species are fully filled

Then, the enlarged lattice is built.

Example

Galois Lattice

(extension x, intension)

(extension y, intension)

Expert

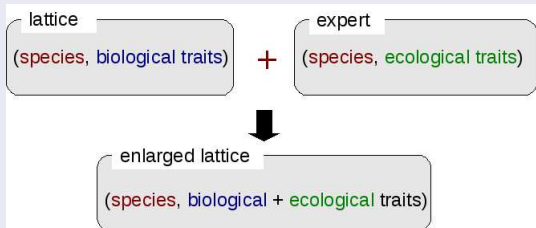
extension x + environment
=> ecological trait A

extension y + environment
=> ecological trait B

	ecological trait A	ecological trait B
object x1	x	
...	x	
object xn	x	
object y1		x
...		x
object yn		x

Validation of the method

General consideration



Theoretical application

- Concept 1 : (JUNA SEFC TYPL, L100-0 H0-100 P100-0-0 V0-100-0-0 D100-0-0) ⇒ mesotrophic to eutrophic water.
- Concept 2 : (CERD PHAA PTCR PTLU PTNA UTRV, H0-100 F0-0-0-100 P100-0-0) ⇒ calm water surfaces

Applicative validation

Concept 1 : (JUNA SEFC TYPL, L100-0 H0-100 P100-0-0 V0-100-0-0 D100-0-0) ⇒
mesotrophic to eutrophic water

Exists in the enlarged lattice and confirms expertise exactly. Its intent includes new attributes : I0-50-50-0 (trophic status) and E0-0-100 (tolerance to sedimentation)

- I0-50-50-0 means species are fairly spread between mesotrophic and eutrophic waters
- E0-0-100 means they have a constant implanting
⇒ Information given by the lattice.

Applicative validation

Concept 2 : (CERD PHAA PTCR PTLU PTNA UTRV, H0-100 F0-0-0-100 P100-0-0) ⇒ calm water surfaces

Exists in the enlarged lattice. 4 of the 6 species share the attributes : A66-33-0 (water level stability) and U100-0-0-0 (resistance to flow)

- A66-33-0 means : 66% live in stable water, 33% in water having fluctuations. Mainly (4 of the 6 species) fit the expertise : live in calm water. PTNA has A40-40-20 (80% live mainly in calm water). PHAA has A0-50-50 attribute which disagrees with the expert.

⇒ Expert indicates the main ecological trait for the group of species. Galois lattice allow to be more precise.

Applicative validation

Concept 2 : (CERD PHAA PTCR PTLU PTNA UTRV, H0-100 F0-0-0-100 P100-0-0) ⇒ calm water surfaces

- U100-0-0-0 indicates 4 species dislike water level variations. PTNA has U0-100-0-0 (bears small variations). PHAA has U0-0-0-100 endures easily important variations.
⇒ Information given by the lattice.

Conclusion

- Definitions of fuzzy many-valued context and histogram scaling to convert complex data into binary context. Keep spreading of the species toward the modalities.
- Proposal of a method based on Galois Lattice to select ecological traits.
- Information given by the expert found again.
- The lattice gives higher accuracy about this expertise and about other ecological traits (not firstly associated by the expert to the concepts).
- The method is reliable.
- Some biologists think ecological traits is an answer to a Water Framework Directive : evaluation of the quality of the whole ecosystem with regard to pressures it endures, because actual evaluation tools are not sufficient.

Future work

- Apply this method to invertebrates. More complex because of several taxonomical degrees and a 314 objects and 1891 histograms dataset.
- Develop algorithm based on Galois connection able to take information about histogram structure and fuzzy consideration into account.



Merci

Si vous avez des questions...