On Mining Temporal Data Using Relational Concept Analysis – An Application to Hydroecological Data

Cristina Nica[1], Agnès Braud[1], Xavier Dolques[1]
Marianne Huchard[2], Florence Le Ber[1]

[1] ICube, University of Strasbourg, CNRS, ENGEES
firstname.lastname@engees.unistra.fr,agnes.braud@unistra.fr
http://icube-bfo.unistra.fr
[2] LIRMM, University of Montpellier, CNRS
huchard@lirmm.fr
https://www.lirmm.fr

1 Experiments and Discussion

RCA is applied by using the RCAExplore[3] tool. The algorithm used during the extraction and selection steps is developed in Java 8 and it allows us to automatically extract closed partially ordered patterns (cpo-patterns).

The Fresqueau[4] project gathers and unifies databases that are linked to waterbodies, as explained in [1]. Five sequential datasets (each dataset concerns only the IBGN[5] biological (Bio) Parameter having the specified quality class) from this project are analysed: $IBGN_{blue}$, $IBGN_{green}$, $IBGN_{yellow}$, $IBGN_{orange}$ and $IBGN_{red}$. The objective is to extract cpo-patterns representing frequent physico-chemical (PhC) trends of watercourses common in many sites. In other words, we are interested in assessing the influence of PhC parameters (e.g. Phosphor (PHOS) and Organic Matter Pollutions (MOOX)) on Bio ones. To this end, the datasets are preprocessed and temporally modelled as described in [2]. Table 1 shows some quantitative statistics regarding the relational analysis and the extraction steps. The relational analysis step relies on the IceBerg algorithm [3], which result is a concept lattice of frequent closed itemsets. A 10% threshold is used only for the input of Bio samples (it corresponds to the main lattice). The choice of this value allows us to focus on the cpo-patterns that describe many sites. The number of extracted cpo-patterns is quite substantial and has to be reduced. To this end, we select relevant cpo-patterns based on their support, richness and the distribution of the associated concept extents.

Henceforward this section refers to the extracted concrete cpo-patterns for $IBGN_{red}$ dataset (63 distinct sites). Figure 1 is a scatter-plot of the distribution IQV of concept extents and the *support*. The diameter of circles is proportional to the richness of the cpo-patterns. By defining two thresholds $\theta_{IQV} = 0.97$ and

[3] http://dolques.free.fr/rcaexplore
[4] http://engees-fresqueau.unistra.fr
[5] Standardised Global Biological Index

Table 1: The results of mining the Fresqueau temporal datasets. `Input Bio` and `PhC` are the initial number of Bio and PhC samples; `Output` is the number of concepts from the main lattice and the lattice of PhC samples; `CPO-patterns` is the number of extracted cpo-patterns.

		RCA				Extraction		
Index	Quality	Input		Output		CPO-patterns		
		Bio	PhC	$\mathcal{L}_{\mathcal{K}bios}$	$\mathcal{L}_{\mathcal{K}phcs}$	Concrete	Abstract	Hybrid
IBGN	blue	108	259	32214	21476	1351	26	30836
	green	127	269	25621	35750	1606	1197	22817
	yellow	103	207	11123	15202	595	457	10070
	orange	108	268	38990	45467	1060	1279	36650
	red	76	169	22232	62040	379	3047	18805

$\theta_{Support} = 20$, the top-24 most well-distributed and most frequent cpo-patterns are selected. The top-24 cpo-patterns describe more or less extensive geographical areas. Consequently, the selected cpo-patterns are ranked by analysing the diameter of circles.

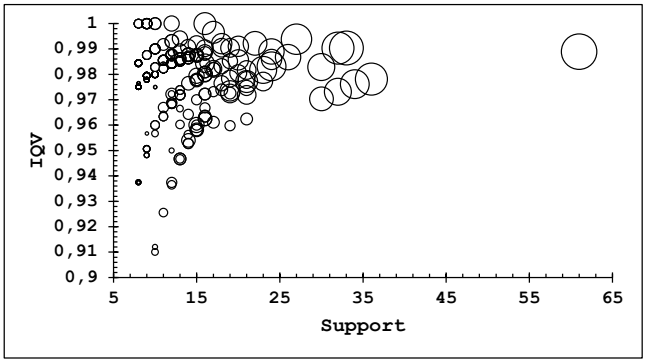

Fig. 1: Concept distribution and support.

The qualitative interpretation of the extracted cpo-patterns was performed by a hydroecologist. The intrinsic general-to-specific ordering of the main lattice concepts allows the expert to begin the analysis from frequent and common PhC trends to particular ones or vice versa. Figure 2 is an excerpt from the main lattice emphasizing the well-known correspondence between `MOOX` quality classes and `IBGN` ones: $\langle(\texttt{MOOX}_{\texttt{red}})\rangle$ covers 32% of the studied area. $\langle(\texttt{PHOS}_{\texttt{red}})\rangle$, which covers more than 33% of the monitored area, is another interesting cpo-pattern since it may highlight the impact of phosphorus pollution on macro-invertebrates (`IBGN`) that is a lesser-known fact. Moreover, thanks to the hierarchical structure of the RCA result (see Fig.2) the hydroecologist easily and quickly identifies the

combination of a *red* PHOS and a *red* MOOX (the po-pattern of CKbios_3006) that has also a stronger impact on macro-invertebrates.

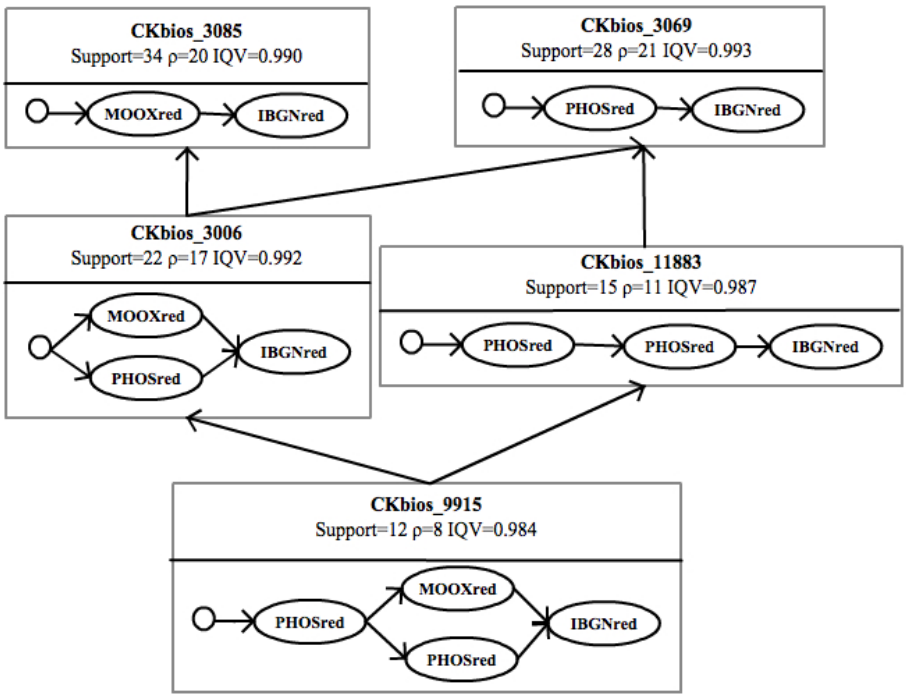

Fig. 2: Hierarchy of cpo-patterns.

References

1. Bimonte, S., Boulil, K., Braud, A., Bringay, S., Cernesson, F., Dolques, X., Fabrègue, M., Grac, C., Lalande, N., Le Ber, F., Teisseire, M.: Un système décisionnel pour l'analyse de la qualité des eaux de rivières. Ingénierie des Systèmes d'Information 20(3), 143–167 (2015)
2. Nica, C., Braud, A., Dolques, X., Huchard, M., Ber, F.L.: L'Analyse Relationnelle de Concepts pour la Fouille de Données Temporelles – Application à l'Étude de Données Hydroécologiques. Revue des Nouvelles Technologies de l'Information Extraction et Gestion des Connaissances, RNTI-E-30, 267–278 (2016)
3. Stumme, G.: Efficient data mining based on formal concept analysis. In: Hameurlain, A., Cicchetti, R., Traunmller, R. (eds.) Database and Expert Systems Applications, pp. 534–546. Springer Berlin Heidelberg (2002)