RCAExplore, a FCA based Tool to Explore Relational Data

Xavier Dolques[1], Agnès Braud[2], Marianne Huchard[3], and Florence Le Ber[1]

(1) ICube, Université de Strasbourg, ENGEES, CNRS
{xavier.dolques, florence.leber}@engees.unistra.fr
(2) ICube, Université de Strasbourg, CNRS
agnes.braud@unistra.fr
(3) LIRMM, Université de Montpellier 2, CNRS
huchard@lirmm.fr

Abstract. Relational Concept Analysis (RCA) is one variant of Formal Concept Analysis for multi-relational dataset exploration. The tool RCAExplore is an implementation of the RCA process where several choices can be made before each iteration: the structure to be used (concept lattice, AOC-Poset, Iceberg lattice), the scaling quantifier (*exist*, *forall*, *contains*, *percentage-quantifiers*, etc.), and the considered formal and relational contexts. RCAExplore was developed during Fresqueau ANR 11 MONU 14 project, in order to explore relational hydroecological data, and has also been used on several datasets in different other domains. The source code and a standalone *jar* application are available online. An integration of the tool is ongoing within a platform for data exploration and knowledge representation.

Keywords: Data exploration · Formal Concept Analysis · Relational Concept Analysis

1 Relational Concept Anaysis (RCA) basics

Relational Concept Analysis (RCA) is an extension of Formal Concept Analysis [5] which considers relational data, formalized within a *relational context family* (RCF), i.e. a set **K** of object-attribute contexts (object categories) and a set **R** of object-object contexts (relations between objects of various categories).

For example, Figure 1 sketches a hydroecological dataset of Fresqueau project. It describes through different relations river stations and taxons living there. Taxons are characterized by life traits (e.g. their locomotion mode). Stations belong to a certain kind of watercourse. Physico-chemical parameters (temperature, oxygen, nitrite, etc.) have been measured on the stations, and their values segmented into two levels. Taxons have been counted from samples and their number segmented into three intervals.

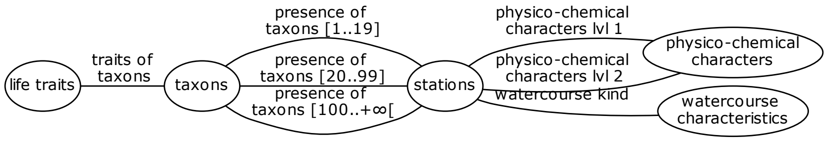

life traits
traits of taxons
taxons
presence of taxons [1..19]
presence of taxons [20..99]
presence of taxons [100..+∞[
stations
physico-chemical characters lvl 1
physico-chemical characters lvl 2
watercourse kind
physico-chemical characters
watercourse characteristics

Fig. 1. A schema of data analysed by RCAExplore, from [4]

Exploring such a relational dataset allows to answer several questions, e.g. *what are the links between life traits of the taxons and values of physico-chemical parameters on river stations?* Answers can take several forms, such as extracting rules that involve the relations, or grouping objects from the different categories (like physico-chemical parameters, or taxons) depending on their attributes and on the objects of another category they are connected with.

RCA can be used for such tasks. First, the user has to build object-attribute contexts, here, `life traits`, `taxons`, `stations`, `physico-chemical characters` and `watercourse characteristics`. Relations between objects are then formalised into seven object-object contexts, e.g. `presence of taxons1-19` goes from `taxons` to `stations`: a taxon is linked to a station, if the sample size for this taxon on this station is between 1 and 19 (see RCAExplore editor open on this relation, in Fig. 3).

The principle of RCA consists in integrating object-object relations as new attributes (called *relational attributes*) in formal contexts thanks to scaling quantifiers, such as the existential (*exist*) and universal strict (*exist+forall*) scaling quantifiers. It produces iteratively a set of concept lattices (one lattice per object category) interconnected through relational information. The concepts in a given lattice group objects according to the shared attributes and to the connections they have with objects of another category. The result is a family of concept lattices where concepts of a lattice are linked to concepts of another lattice, as sketched out in Fig. 2: a concept C_s reveal a group of stations (1) where at least one physico-chemical value from a group of physico-chemical characteristics (concept C_p) has been measured with level 2, (2) being in a kind of watercourse from a watercourse characteristics group (concept C_w), (3) hosting more than 60% of numerous (100+) taxon types from a group (concept C_t) whose elements have at least one trait from a group of life traits (concept C_l).

The process is as follows: first, lattices are built on the object-attribute contexts K_i; second, object-attribute contexts are extended with new attributes built using the object-object relations and the concept sets of the lattices; third, new lattices are built on the extended contexts. For instance let $\mathbf{K} = \{K_1, K_2\}$, $\mathbf{R} = \{R\}$, where $K_1 = (G_1, M_1, I_1)$, $K_2 = (G_2, M_2, I_2)$ and $R = (G_1, G_2, r)$. Let $C = (X, Y)$ be a concept of the K_2 lattice (i.e. $X \subseteq G_2$); $\exists$ is the existential quantifier. A relational attribute $\exists r(C)$ is added to the K_1 context, and is owned by an object $g \in G_1$ if $r(g) \cap X \neq \emptyset$. A whole construction process consists in building a finite sequence of contexts and concept lattices based on $(\mathbf{K}, \mathbf{R})$ and

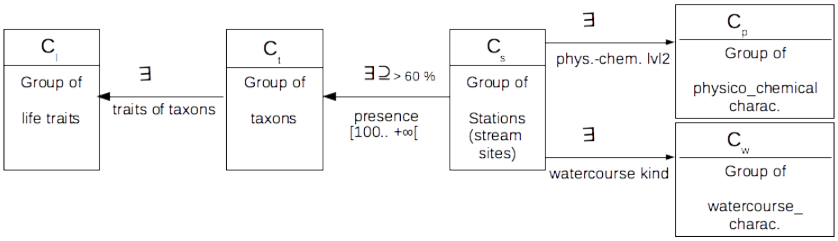
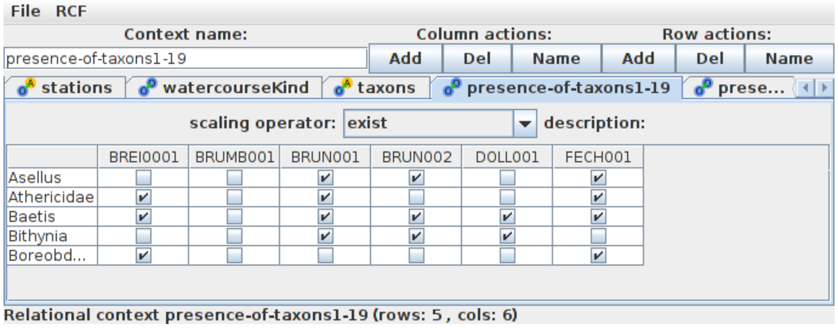


Fig. 2. Illustration of the results of RCA on the dataset described in Fig. 1: concepts are linked by relational attributes, ∃, and $\exists\supseteq_{\geq 60\%}$ are scaling quantifiers

chosen scaling quantifiers. The last sequence is obtained when the fixed point is reached. Details about RCA are given in [6].

2 RCAExplore

RCAExplore[1] is a reference implementation of RCA offering new ways of dealing with relations. RCAExplore is developed in Java. It proposes a relational context family editor (see Fig. 3, highlighting an object-object context linking `taxons` and `stations`), an interactive concept lattice family generator (see Fig. 4), and a concept lattice family browser. The tool is developed under the LGPL license.



Context name: presence-of-taxons1-19

scaling operator: exist description:

	BREI0001	BRUMB001	BRUN001	BRUN002	DOLL001	FECH001
Asellus	☐	☐	☑	☑	☐	☑
Athericidae	☑	☐	☑	☐	☐	☑
Baetis	☑	☐	☑	☑	☑	☑
Bithynia	☐	☐	☑	☑	☑	☐
Boreobd...	☑	☐	☐	☐	☐	☑

Relational context presence-of-taxons1-19 (rows: 5 , cols: 6)

Fig. 3. Editing a relational context family; files are saved in a readable format `.rcft`

The novelty behind RCAExplore, with respect to other RCA tools such as Galicia[2], is to provide, besides an automatic mode, a manual mode where it is possible to modify the data considered before each new iteration step of the RCA process. At each step of the process, the user can choose (Fig. 4):

[1] http://dataqual.engees.unistra.fr/logiciels/rcaExplore
[2] http://www.iro.umontreal.ca/∼galicia/

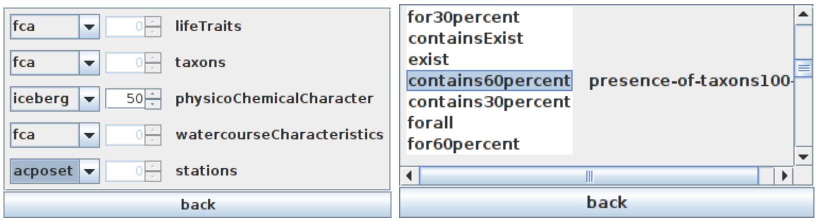
- the conceptual structure (concept lattice, AOC-poset, Iceberg lattice),
- the scaling quantifier (in $\{\exists, \exists\forall, \exists\supseteq, \exists\forall_{\geq n\%}, \exists\supseteq_{\geq n\%}\}$, $n = 30$ and $n = 60$),
- the considered contexts (among the available object-attribute and object-object contexts),
- if they want the simplified or the full intents and extents, and other options.

The tool checks whether the choices made by the user are consistent. For example, a user cannot choose at step n a relational context with the target formal context K, if K has not been chosen at step $n-1$. Files are generated at each step, that contain all the conceptual structures of this step in a format (`.svg` and `.dot`, from Graphviz[3]) which is readable and exploitable by other programs, and it is possible to also see the extended contexts in LaTeX and HTML. A trace file records the choices made during the process, and can be used to replay the same configuration in automatic mode, namely without again manually selecting structures, quantifiers or contexts. A visualization option also allows to see graphically, at the end of the process, the conceptual structures that have been generated at all steps.

The concept lattice family browser allows, from a given step number, a given context and a given concept name, to look at the parents, children, intent, simplified intent and extent of a concept. If two attributes are selected in the simplified intent and the intent respectively, the corresponding implication rule is displayed.

3 Application examples

As previously introduced, RCAExplore has been primarily used to explore hydroecological data in the framework of Fresqueau ANR 11 MONU 14 project. The question was to link the characteristics (traits) of taxons to the physico-chemical state of a river station where they live. Data have been explored according to different paths: focusing on river stations and trying to characterize them both with taxon traits and physico-chemical parameters, focusing on specific relations (e.g. presence of rare taxons in river stations) by varying and combining



fca	0	lifeTraits
fca	0	taxons
iceberg	50	physicoChemicalCharacter
fca	0	watercourseCharacteristics
acposet	0	stations

for30percent
containsExist
exist
contains60percent presence-of-taxons100-
contains30percent
forall
for60percent

Fig. 4. Choosing the conceptual structure and scaling operators to be used

[3] https://www.graphviz.org/

the scaling operators, etc. Rules have been computed [3], e.g. a straightforward rule[4], could be (meaning that if a station hosts many of numerous taxon types that appreciate slow current, then the river has calm water):

$$\begin{array}{ll} \exists\supseteq_{\geq 60\%} & \text{presence}_{[100,..+\infty[}(\exists\text{traits\_of\_taxons}(\texttt{appreciate slow current})) \\ \rightarrow & \exists\text{watercourse\_kind}(\texttt{calm water}) \end{array}$$

The features of RCAExplore have been useful to support FCA and several research experiences, to mention just a few: effect of varying the scaling quantifiers on a subset of the Fresqueau dataset is shown in [1], assessing the relevance of using AOC-posets during gradual class model refactoring has been shown in [7] on industrial UML and Java class models. In [2], the authors use RCAExplore to assess the scalability of RCA on connected product comparison matrices. RCAExplore has also been used to analyse spatio-temporal data characterizing river networks [8] and is currently used to teach FCA and RCA at Montpellier and Strasbourg Universities. It is under integration in the COGUI platform[5] under the guidance of Alain Gutierrez.

References

1. Braud, A., Dolques, X., Huchard, M., Le Ber, F.: Generalization effect of quantifiers in a classification based on relational concept analysis. Knowl.-Based Syst. **160**, 119–135 (2018)
2. Carbonnel, J., Huchard, M., Gutierrez, A.: Variability representation in product lines using concept lattices: Feasibility study with descriptions from wikipedia's product comparison matrices. In: Proc. of the Int. Ws. on Formal Concept Analysis and Applications, FCA&A 2015, co-located with (ICFCA 2015). pp. 93–108 (2015)
3. Dolques, X., Le Ber, F., Huchard, M., Grac, C.: Performance-friendly rule extraction in large water data-sets with AOC posets and relational concept analysis. Int. J. General Systems **45**(2), 187–210 (2016)
4. Dolques, X., Le Ber, F., Huchard, M., Nebut, C.: Relational Concept Analysis for Relational Data Exploration. Adv. in Know. Disc. and Manag. **5**, 55–77 (2015)
5. Ganter, B., Wille, R.: Formal Concept Analysis: Mathematical Foundations. Springer Verlag (1999)
6. Hacene, M.R., Huchard, M., Napoli, A., Valtchev, P.: Relational concept analysis: mining concept lattices from multi-relational data. Ann. Math. Artif. Intell. **67**(1), 81–108 (2013)
7. Miralles, A., Molla, G., Huchard, M., Nebut, C., Deruelle, L., Derras, M.: Class Model Normalization - Outperforming Formal Concept Analysis Approaches with AOC-posets. In: Proc. of the 12th Int. Conf. on Concept Lattices and Their Applications (CLA'15). pp. 111–122 (2015)
8. Nica, C., Braud, A., Le Ber, F.: Exploring Heterogeneous Sequential Data on River Networks with Relational Concept Analysis. In: Proc. of the 23rd Int. Conf. on Conceptual Structures (ICCS'18). pp. 152–166. LNAI 10872 (2018)

[4] Such straightforward rules are useful to understand and to build confidence on other, more informative, rules.

[5] https://www.lirmm.fr/cogui/3/