Adjusting the Exploration Flow in Relational Concept Analysis

Amirouche Ouzerdine, Agnès Braud, Xavier Dolques, Marianne Huchard, Florence Le Ber

Adjusting the Exploration Flow in Relational Concept Analysis

An Experience on a Watercourse Quality Dataset

Amirouche Ouzerdine, Agnès Braud, Xavier Dolques, Marianne Huchard, and Florence Le Ber

Abstract In this paper, we focus on the exploration of multi-relational datasets, and the various ways they can be analyzed using Relational Concept Analysis (RCA), an extension of Formal Concept Analysis (FCA). RCA uses several scaling operators that make the process highly tunable, allowing a high flexibility in the exploration and in the results. In return, the multiplicity of choices that can be made when performing an analysis task potentially overwhelms the expert. We thus propose three overlays for helping users control and foresee the results of their choices. Our proposition is exemplified on a dataset about the hydro-ecological state of watercourses.

Keywords Multi-relational dataset · Relational concept analysis · Formal concept analysis · Relational data exploration

1 Introduction

Multi-relational datasets are based on a schema (data model), where entities (objects) of several categories are described by characteristics (attributes, fields) and where relations link objects from two categories (possibly from the same one). Experts of the domains in which these data are collected are interested in exploiting them

A. Ouzerdine · M. Huchard (✉)
LIRMM, Université de Montpellier, CNRS, Montpellier, France
e-mail: marianne.huchard@lirmm.fr

A. Ouzerdine
e-mail: labib23dz@hotmail.com

A. Braud
Université de Strasbourg, CNRS, ICube UMR 7537, 67000 Strasbourg, France
e-mail: agnes.braud@unistra.fr

X. Dolques · F. Le Ber
Université de Strasbourg, CNRS, ENGEES, ICube UMR 7537, 67000 Strasbourg, France
e-mail: xavier.dolques@engees.unistra.fr

F. Le Ber
e-mail: florence.leber@engees.unistra.fr

R. Jaziri et al. (eds.), *Advances in Knowledge Discovery and Management*,
Studies in Computational Intelligence 1004,
https://doi.org/10.1007/978-3-030-90287-2_9

through multiple tasks: browsing or exploration, querying, extraction of knowledge patterns, or classification. By classification, we mean grouping, hierarchically organizing groups by generalization, and explicitly describing sets of similar objects by characteristics. This paper focuses on *exploration* and *knowledge extraction* in hierarchical by-generalization classifications built on top of the dataset. These tasks are indeed facilitated thanks to: (1) a classification of extracted knowledge patterns rather than a flat set of patterns, and (2) a support for exploring groups of similar objects connected through inter-group links, rather than exploring individual objects connected through inter-individual links.

Formal Concept Analysis (FCA, Ganter and Wille 1999) and its extensions bring methods that contribute to many data exploitation tasks. FCA based on graph representations (Liquière and Sallantin 1998; Kötters 2013; Ferré 2015), logical representations (Ferré et al. 2005) and multiple binary relations such as Relational Concept Analysis (Hacene et al. 2013), are extensions that can apply, in different manners, to multi-relational datasets.

Relational Concept Analysis (RCA) has been specifically designed for data exploration tasks. It iteratively builds a set of interconnected classifications, and it can be used to extract object groups, knowledge patterns and implication rules involving the inter-object links. It has been successfully used for analyzing datasets in different domains such as software engineering (Dolques et al. 2012), ontology engineering (Bendaoud et al. 2008; Rouane-Hacene et al. 2011), Web services (Azmeh et al. 2011), or recently linked data (Atencia et al. 2020).

One main feature of RCA is the process of building abstractions of inter-objects links over object groups (concepts) through quantifiers inspired by description logics constructors, such as *at least one / only / all / at least 30% / etc.* These link abstractions may group objects that have *at least one / only / all / at least 30% / etc.* (of) their outgoing links for a specific relation entering into another identified group of objects. This allows the abstraction behind an object group to be propagated through chained inter-objects links. The step-by-step and exploratory nature of RCA, where the relations and the quantifiers considered at each step can be chosen, makes the process highly tunable.

On the one hand, these tuning possibilities make the resulting classification highly expressive with descriptions using quantifiers beyond the usual universal and existential ones. On the other hand, the multiplicity of choices needed to perform an analysis task can be overwhelming. In Dolques et al. (2015), an adaptation of RCA is proposed to explore relations gradually by configuring one step at a time instead of having to manage the whole process at once. Nevertheless, as far as we know, no tool has been provided to guide the users in their choices and make the process more intuitive.

In this paper, we propose to introduce pieces of knowledge in the RCA process to make the analysis easier, both some knowledge on the data that is given by the users to constrain the process before its start, and some knowledge built on the fly by the users, based on the information extracted from the process, and that help them decide how to tune this process. This takes the form of three overlays over the RCA process: the first overlay consists in expressing constraints on the scaling quantifiers

to keep their choice consistent; the second overlay consists in translating high-level (often schema-level) query patterns, that are difficult to formalize by experts, into expressions of a controlled language; the third overlay consists in giving quantitative metrics on the concept lattices to be built and on particular relational implication rules in order to help experts tune the analysis. The term of "knowledge" is to be understood as defined by Zeleny (2013), "a purposeful coordination of actions", in our case, a coordination of actions to control the data exploration process. Data is the input of the process, information is the output. We apply our proposal on a watercourse quality dataset from the FRESQUEAU project[1] in order to analyze the relations between the physico-chemical state of a river site and the characteristics of taxons (macroinvertebrates) living there.

The paper is organized as follows. In Sect. 2, we deepen the kind of relational data exploration we aim to achieve and why it can be difficult for experts to conduct such analyses. Proposed approaches of the literature, that address the problem of exploring multi-relational datasets through by-generalization or by-aggregation organizations are presented in Sect. 3. Section 4 briefly introduces RCA for our context: its input, its main principles and its outputs. Section 5 describes the overlays we have developed and Sect. 6 presents results on the FRESQUEAU data. We conclude the paper and give some work perspectives in Sect. 7.

2 Exploring Multi-relational Datasets

Many data are inherently relational, with different types of relations. This motivates the development of methods that aim for example to extract relational patterns, or relational association rules, to induce relational decision trees, or to make clustering with relational distance-based approaches (Džeroski 2003). In FCA context, relational data are taken into account through graph-based representations (Liquière and Sallantin 1998; Kötters 2013; Ferré 2015), logical expressions (Ferré et al. 2005), or multiple binary relations, as in RCA (Hacene et al. 2013).

When the experts have vague knowledge on the data and when their queries are general, an exploratory approach can be suitable (Wildemuth and Freund 2012; Palagi et al. 2017). The data exploration can be fully free, or it can be guided by general questions, often at the level of the data schema (concepts and relations).

We use as illustration a typical exploration case for an hydro-ecologist who studies the effect of the physico-chemical state of a watercourse on the characteristics (life traits) of the taxons (animals or plants) living there. An excerpt of the data model is shown in Fig. 1, where water samples have a certain abundance of taxons (identifiers of groups of living beings, here macroinvertebrates, organized into genus and families); those taxons having some affinity with some (modalities of) life traits (e.g. maximal size; aquatic stage: egg, larva, nymph; breathing mode; locomotion mode). Water samples are also described by measures on physico-chemical (PC) parame-

[1] http://engees-fresqueau.unistra.fr/.

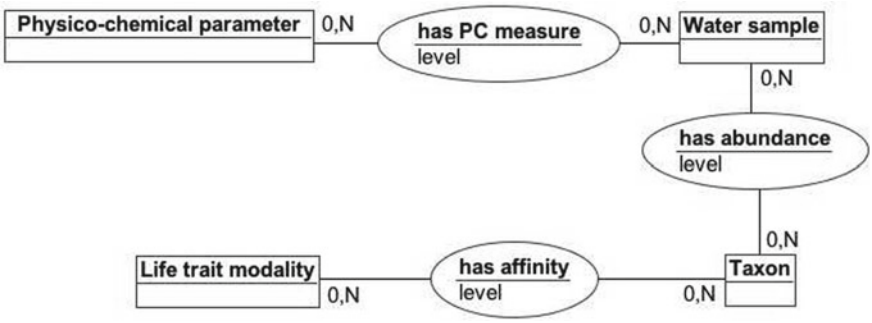

Fig. 1 Excerpt of the data model of the FRESQUEAU project (Bimonte et al. 2015)

ters (e.g. nitrites, minerals, organic matters, temperature) organized into categories depending on their nature. For each relation, the level corresponds to 5 degrees of intensity depending on the number of individuals of a taxon in the water sample for *abundance*, the part of a taxon population showing a life trait modality for *affinity* and on the value of the measure done in the water sample for the *PC parameter measures*. For the analysis purpose, each relation between categories of objects is divided into 5 relations corresponding to those levels.

One main general question of the project experts is: *what are the links between life traits of the taxons and values of physico-chemical parameters?* Exploring the dataset in order to answer the question may take several forms, such as extracting rules that involve the relations, or grouping objects from the different categories (like water samples, or taxons) depending on their attributes and on the objects of another category they are connected with. For example, experts may be interested by the results of the following query: *find groups of water samples that have (1) a certain abundance level for a group of taxons, having themselves a common group of life traits with a certain level of affinity and that have (2) at a certain level, physico-chemical parameters in a certain group*. Results may reveal for example: a group of water samples containing taxons with a long life, and containing much organic matters; a group of water samples with a high level of mineral material, and containing taxons that use crawling as a locomotion mode. If we look closer at it, the general question and the extracted groups can be refined in many directions, such as (terms in bold denote the variable points in a question):

- *find groups of water samples that have (1) a* ***certain*** *abundance level for a group of taxons (and* ***only*** *taxons of this group), having themselves* ***at least 70%*** *of life traits in a common group of life traits with a* ***certain*** *level of affinity, and that have (2) at a* ***certain*** *concentration level,* ***at least one*** *physico-chemical parameter in a certain group.*
- *find groups of water samples that have (1) a* ***certain*** *abundance level for a group of taxons (and* ***more than 60%*** *of the taxons of each water sample are in this group), having themselves* ***only*** *life traits in a common group of life traits with a*

certain *affinity level, and that have (2) at a* ***certain*** *concentration level,* ***all the*** *physico-chemical parameters in a certain group*.

On the one hand, reformulating the general query in some of the possible directions is important for experts because they need precise answers and they also can change the focus of their analysis. On the other hand, the refinement can be done in many directions, and experts will probably get lost relatively quickly. Besides, choosing some refinements compared to others may lead to a too restrictive or too large set of results. Lastly, the results are sets of connected groups of objects that respect a certain question pattern. These groups may be many and they potentially specialize one another, as shown in the following.

3 By-generalization or By-aggregation Approaches for Multi-relational Dataset Exploration

Graphs are natural settings for representing relational data. Conceptual graphs have been introduced to represent semantic information (Sowa 2008): they allow to perform queries and generalization/specialization operations based on hierarchies of concepts and relations. These last years, many methods of graph mining have been proposed, and more specifically in multi-graphs (Ingalalli et al. 2018). In the domain of visualization, multi-relational datasets are represented as multilayer graphs. For example, VERTIGo (Cuenca et al. 2018) is a visualization-based data mining system that groups links and objects during navigation to present query results at different levels of details and suggest new query extensions.

Inductive Logic Programming (ILP) was initially concerned with learning logic programs, and ILP techniques have then been applied in relational data mining. In ILP, learning is performed directly in the first-order logic setting, so that the search space is intractable when data are numerous. Propositionalisation was proposed as a mean to reduce this complexity (Muggleton and Raedt 1994; Lachiche 2010). In Dolques et al. (2014), a comparison between propositionalisation and RCA has been conducted. Besides, relational data have been transformed into logical formulae within the framework of logical concept analysis described in Ferré et al. (2005).

Data Warehouses (DWs) are databases dedicated to the integration and storage of large volumes of data to support the decision processes of organizations (Inmon 2005). DWs store decisional data at the finest granularity level and organize them to facilitate analysis and aggregation. On-Line Analytical Processing (OLAP) tools allow to build multidimensional data structures having different granularities, called data cubes, by aggregating DW data, and provide users with operators for rapid exploration of these data cubes (Kimball and Ross 2002). The dimensions are organized into hierarchies of aggregation (or generality) levels to allow viewing analysis indicators at different granularities. Such cubes have been implemented to aggregate and navigate FRESQUEAU data (Boulil et al. 2014).

Formal Concept Analysis approached the multi-relational datasets through several perspectives. Some approaches extract and classify graph patterns that connect objects or object groups (Liquière and Sallantin 1998; Prediger and Wille 1999; Ganter and Kuznetsov 2001). Besides, K. E. Wolff has introduced Relational Semantic Systems: the data model is represented through a conceptual graph, while the relational knowledge is represented through *object traces* and *relation concept traces* in *trace diagrams* (Wolff 2009). Tuples of boolean factors are extracted from various tables thanks to an extended version of the Boolean Factor Analysis (Krmelova and Trnecka 2013). An n-ary relation may be in many concrete cases considered as an aggregation of several relations of lower arity. FCA has thus been generalized into Triadic Concept Analysis, that considers a ternary relation including objects, attributes and conditions. This yields *triadic concepts* that are organized in a complete *trilattice*. This framework has been generalized to n-adic contexts (n-ary relations) in Polyadic Concept Analysis (Voutsadakis 2002).

Although the previous FCA approaches give relevant and complementary views on multi-dimensional datasets, they do not specifically focus on dataset exploration. A step towards data exploration is made through the definition of queries. In the existential case, this has been proposed by J. Kötters with relational, windowed structures (Kötters 2013), and by S. Ferré in Graph-FCA (Ferré 2015). A combined approach takes advantage of OLAP cubes to structure sets of concepts and OLAP operations to support multi-dimensional navigation (Ferré et al. 2012). In the spirit of conceptual navigation (Carpineto and Romano 2004; Wray and Eklund 2011; Dunaiski et al. 2017), Abstract Conceptual Navigation (Ferré 2014) introduces a high-level view on exploration in the context of FCA, which gives guidelines and inspiration to define exploratory approaches.

Relational Concept Analysis (RCA) (Hacene et al. 2013) more specifically focuses on exploring datasets by highlighting the object categories (each being encoded in a formal context), and by connecting objects from the same category or from different categories through relational contexts. The result is a set of interconnected concept lattices (one lattice per object category). Connections between the lattices are made through *relational attributes* which capture groups of similar individual links between objects. The relational attributes are built using operators inspired by description logics. By nature, RCA is iterative, as concepts formed on one object category emerge and are propagated step-by-step through the constructed relational attributes. Thus, immediate relational abstractions come first. Combining RCA and Graph-FCA to help the RCA results interpretation has been studied in Ferré and Cellier (2018). In the next section, we introduce basics on RCA.

4 Data Exploration with Relational Concept Analysis

Relational Concept Analysis (Hacene et al. 2013) goes beyond Formal Concept Analysis (Ganter and Wille 1999) by considering a multi-relational dataset. With respect to the previously mentioned methods, it adds a set of operators and an iter-

ative approach that allows to follow information propagation through the various relations, helps the comprehension of patterns, rules or clusters formation process, and facilitates user (even abductive) reasoning. A relational dataset is represented by a Relational Context Family (RCF), composed of formal contexts and relations. Each formal context (also called object-attribute context) represents a set of objects of a given category by their attributes. The relations (also called object-object contexts) connect the objects of the different categories (or of the same category).

Definition 1 (*Relational Context Family (RCF)*) A Relational Context Family is a $(\mathbf{K}, \mathbf{R})$ pair where:

- $\mathbf{K} = \{\mathscr{K}_i\}_{i=1,\ldots,n}$ is a set of $\mathscr{K}_i = (G_i, M_i, I_i)$ object-attribute contexts, where G_i is a set of objects, M_i is a set of attributes and $I_i \subseteq G_i \times M_i$; and
- $\mathbf{R} = \{r_j\}_{j=1,\ldots,p}$ is a set of r_j relations (object-object contexts) where $r_j \subseteq G_k \times G_l$ for some $k, l \in \{1, \ldots, n\}$.

A simple example of RCF inspired by our hydro-ecological application domain is shown in Table 1 (*has_abundance*, is here binary for the sake of illustration). The `Taxon` formal context introduces:

- the taxons *Aeschnidae* (`Aes.`), *Agabus* (`Agb.`), *Agraylea* (`Aga.`), *Agriotypus* (`Agi.`), *Ancylus* (`Anc.`), *Anisus* (`Ani.`), *Anodonta* (`Ano.`), *Anthomyiidae* (`Ant.`).
- five attributes describing their micro-habitats (boulders, gravel, sand, macrophytes, organic detritus/litter).

The `WaterSample` formal context describes 8 water samples by their flow characteristic (torrent, calm water) and chemical components. The `has_abundance` object-object context connects the 8 water samples to the taxons that have been found into during a sampling campaign.

Formal Concept Analysis can be applied to the formal contexts `Taxon` and `WaterSample` to form hierarchies of object groups sharing common attributes. These groups are called concepts and are more precisely defined as follows.

Definition 2 (*Formal concept*) Given an object-attribute context $K = (G, M, I)$, a *concept* maps a maximal set of objects with the maximal set of attributes they share, yielding a set pair $C = (Extent(C), Intent(C))$ such that:

- $Extent(C) = \{g \in G | \forall m \in Intent(C), (g, m) \in I\}$ is the extent of the concept (objects covered by the concept).
- $Intent(C) = \{m \in M | \forall g \in Extent(C), (g, m) \in I\}$ is the intent of the concept (shared attributes).

The formal concepts are ordered through a specialization/generalization order, denoted by $\preceq_C$, based on the set-inclusion order. Given two formal concepts $C_1 = (E_1, I_1)$ and $C_2 = (E_2, I_2)$, $C_2 \preceq_C C_1$ if and only if $E_2 \subseteq E_1$ (and equivalently $I_1 \subseteq I_2$). C_2 is a specialization (i.e., subconcept) of C_1. C_1 is a generalization (i.e., superconcept) of C_2. C_2 intent inherits the attributes from C_1 intent, while C_1 extent

Table 1 RCF composed of the formal contexts `Taxon`, `WaterSample`, and the object-object context `has_abundance`

Taxon	boulders	gravel	sand	macrophytes	orgLitter
Aes					×
Agb				×	×
Aga				×	×
Agi	×	×	×		×
Anc	×	×			
Ani	×	×	×		×
Ano				×	×
Ant	×	×			

WaterSample	torrent	calmWater	NH4	SO4	Ca	Mg	C3H8NO5P
ws1	×		×				
ws2	×		×				
ws3		×		×			
ws4		×		×			
ws5		×		×			×
ws6		×		×			×
ws7		×		×		×	
ws8		×		×	×		

has_abundance	Aes	Agb	Aga	Agi	Anc	Ani	Ano	Ant
ws1					×			×
ws2					×			×
ws3	×	×		×		×		×
ws4	×	×		×		×		×
ws5				×		×	×	
ws6				×	×	×		
ws7		×	×					×
ws8		×					×	×

inherits the objects from C_2 extent. The set of all concepts of K, ordered by $\preceq_C$, is provided with a lattice structure, and is called the concept lattice of K.

Figure 2 shows the concept lattices associated with the formal contexts of water samples (left-hand side) and taxons (right-hand side). The lattice formed on water samples highlights the group of water samples collected in calm waters ($C\_WaterSample\_5$) versus the group of water samples collected in torrents ($C\_WaterSample\_4$). The water samples collected in calm waters are then separated in three subgroups depending on the presence of glyphosate (`C3H8NO5P`, in $C\_WaterSample\_3$), calcium (`Ca`, in $C\_WaterSample\_1$) or magnesium (`Mg`, in $C\_WaterSample\_2$).

In RCA, relational attributes are introduced to complete the initial formal contexts to take into account relational information. A relation $r_j \subseteq G_k \times G_l$ will be used to build relational attributes of K_k by using relations between objects of G_k and concepts built over objects of G_l. Figure 3 illustrates relational attributes with a few examples, on the set of water samples composed of $ws3$, $ws6$, $ws7$ and $ws8$. A relational

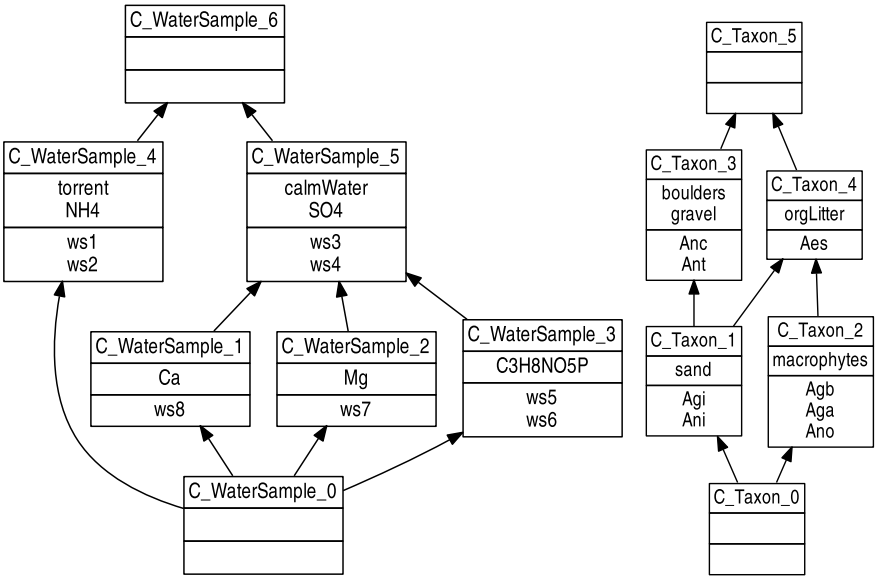

Fig. 2 Concept lattices for formal contexts `WaterSample` and `Taxon`. A concept is shown as a three-parts box. The upper part is its identifier; the middle part contains the intent deprived of the top-down inherited attributes (it contains "introduced" attributes only); the bottom part contains the extent deprived of the bottom-up inherited objects (it contains "introduced" objects only)

attribute is composed of a scaling quantifier, the name of the relation, and the target concept. For example:

- Relational attribute $\exists has\_abundance(Concept\_Taxon\_2)$ is associated with water samples $ws3$, $ws7$ and $ws8$ because they have **at least one** *has_abundance* link to a taxon of the extent of $Concept\_Taxon\_2$.
- Relational attribute $\exists\forall has\_abundance(Concept\_Taxon\_3)$ is associated with water sample $ws6$, because it has **at least one** *has_abundance* link and such links are **only** directed to taxons of the extent of $Concept\_Taxon\_3$.
- Relational attribute $\exists\forall_{\geq 60\%} has\_abundance(Concept\_Taxon\_2)$ is associated with water samples $ws7$ and $ws8$ because they have **at least one** and **at least 60%** of their *has_abundance* links to taxons of the extent of $Concept\_Taxon\_2$.
- Relational attribute $\exists\supseteq has\_abundance(Concept\_Taxon\_1)$ is associated with water samples $ws3$ and $ws6$ because they have **at least one** and **all the taxons** of the extent of $Concept\_Taxon\_1$ through *has_abundance* links.

Using such relational information leads to extend formal contexts with the relational attributes and build new concept lattices. For example, left-hand side of Fig. 4 (resp. right-hand side) shows the concept lattice associated with the formal context `WaterSample` extended with all possible relational attributes composed with scaling quantifier $\exists\forall$ (resp. $\exists\forall_{\geq 60\%}$) and concepts of the `Taxon` concept lattice of Fig. 2. Figure 4 also highlights a generality relation between the scaling quantifiers. In our

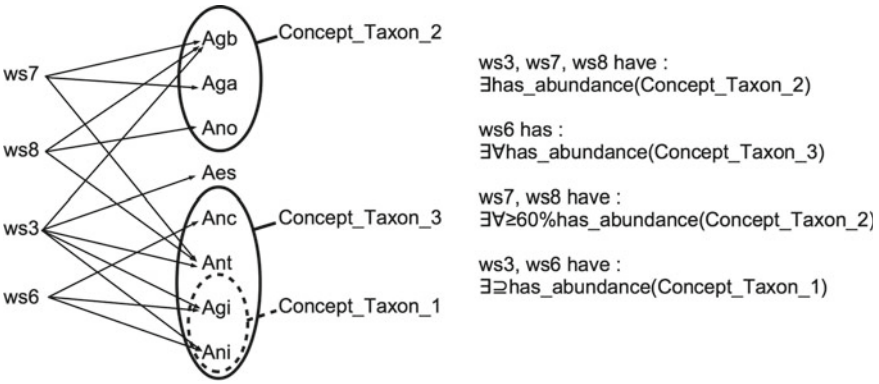

Fig. 3 Relational attributes built from `Taxon` concepts and the relation *has_abundance* between `WaterSample` and `Taxon`

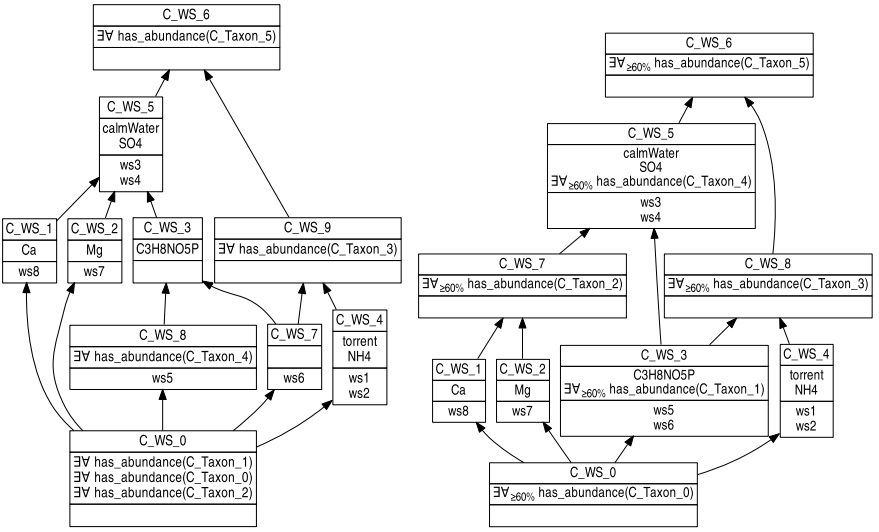

Fig. 4 Concept lattices for `WaterSample` (WS) with scaling quantifiers $\exists\forall$ (LHS) and $\exists\forall_{\geq 60\%}$ (RHS)

example, $\exists\forall$ is more general than $\exists\forall_{\geq 60\%}$ (denoted as $\exists\forall \preceq_S \exists\forall_{\geq 60\%}$), with the consequence that if an object owns a relational attribute formed with $\exists\forall$, it also owns its equivalent (same relation/same concept) formed with $\exists\forall_{\geq 60\%}$, and there is a form of projection between the relational attribute introducers in the left-hand side lattice to these of the right-hand side lattice (Braud et al. 2018).

Let us note that the hydro-ecologist queries look like statements that can be written using query languages like SQL, however in the above examples the groups are formed during the RCA process, helping inherently to deliver the results in an orga-

nized form. For example, if *finding calm water samples* is the expert query, the lattice from the left-hand side of Fig. 2 organizes the answers with $C\_WaterSample\_5$ and its subconcepts. The expert learns which answers correspond to its query with the least possible amount of additional characteristics, and the more she/he goes down the lattice, the more characteristics are added to the groups. She/he also learns which water samples are among the "equivalent" answers (answers that have exactly the same characteristics), or which characteristics appear together. All these informations help her/him to navigate among the possible answers to the query. Using RCA, these lessons learned extend to the relational information. For example, the concept lattice of Fig. 4 (right-hand side) highlights the fact that water samples from calm water (C_WS_5) have a significant proportion of their taxons among those which appreciate organic litter. It more specifically represents the answers to the general query: "find groups of water samples that have more than 60% of their taxons in a certain taxon group". C_WS_5 and its subconcepts show the organization of the answers to the query *finding calm water samples* or alternatively *finding water samples that have more than 60% of their taxons in the micro-habitat organic litter*. This specific feature of FCA and RCA brings data structuring to the expert attention and fosters the navigation within the answers, and within the dataset.

In the general case, the data model can be cyclic: for example, we could have the reverse relation $is\_abundant\_in$ from taxon to water samples. In this case, once concepts of water samples are built, a new concept lattice of taxons can be built thanks to a chosen scaling quantifier and the resulting relational attributes formed on $is\_abundant\_in$. This entails an iterative process which converges after a number of steps depending on the dataset. Besides, the initial definition of the RCA process considers that a single scaling quantifier is associated with a given relation all along the iteration process. But variants have been defined and used for specific usages (Dolques et al. 2015; Braud et al. 2018).

5 Guiding Tools for RCA

The tool RCAexplore[2] allows a variety of RCA usages: changing at each step the scaling quantifiers, the considered formal contexts and relations and the set of concepts that are computed. This variety of usages has its counterpart which is the difficulty of choosing the right parameters for a given question.

To overcome these difficulties, we have designed three overlays, that are added to the general RCA process, as outlined in Fig. 5. In a first step (step 1), a data model (objects, attributes and relations) is chosen; based on this model, the first overlay allows the user to put constraints on relations (Sect. 5.1). Step 2 focuses on data processing and formatting, in order to build the RCF; then FCA is applied on the object-attribute contexts (step 3). The choice of which scaling operator to apply to which relations is done in step 4 with the help of the second and third overlays: they

[2] http://dataqual.engees.unistra.fr/logiciels/rcaExplore.

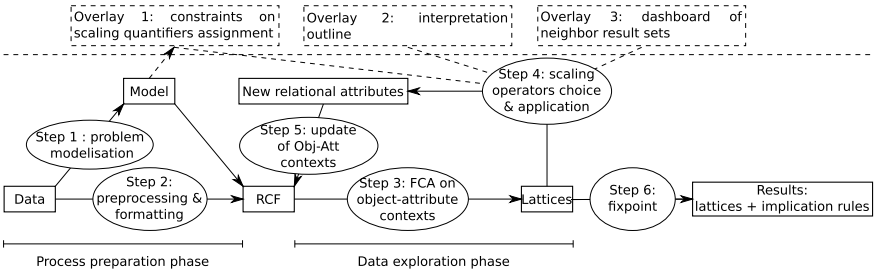

Fig. 5 RCAExplore process and its overlays

aim to guide the user through scaling quantifiers assignment, with an interpretation outline (Sect. 5.2) and information relative to neighbor result sets (Sect. 5.3). The application of scaling operators leads to new relational attributes that complete the object-attribute contexts (step 5); then step 3, step 4 and step 5 are again applied, until a fixpoint is reached (step 6). The obtained results are concept lattices, or other conceptual structures, as well as different extracted information, such as implication rules. These overlays have been introduced for the FRESQUEAU dataset, jointly with hydro-ecologists, which explains the applied perspective, but our experience with RCA makes us think that they have a wider interest for other multi-relational datasets in other domains. A short demo is available.[3]

5.1 Constraints on Relations

RCAExplore offers the possibility to choose among several quantifiers on relations, but sometimes, some relations are semantically connected and the quantifiers that are associated with them have thus to be consistent. For example, in the FRESQUEAU project, each general relation (e.g. *has_abundance*) is represented with several relations to capture the notion of levels of Fig. 1, e.g. the five relations $has\_an\_abundance\_of\_level\_i$, each one corresponding to a level between 1 and 5. In this case, if several *has_abundance* relations are selected together, it may be consistent to apply to them the same quantifier. Relations are thus gathered into equivalence classes: relations in the same equivalence class are considered in the same way along the process, i.e. they are all given the same scaling quantifier at each step. Nevertheless, the scaling quantifier for a class can be different from one step to another.

This information is encoded in a *json* file which is analyzed each time a scaling quantifier is associated with a relation through the user interface, in order to propagate the constraint to the other relations that are in the same class as the chosen one. The users can accept or change the system propositions. For example, the *json* file shown

[3] http://dataqual.engees.unistra.fr/logiciels/rcaExplore/rcaexplore2019.mp4.

```
{
  "Equality" : {
    "PC parameter" : ["has_for_level_1_PC_measure","has_for_level_2_PC_measure",
    "has_for_level_3_PC_measure","has_for_level_4_PC_measure",
    "has_for_level_5_PC_measure"],

    "affinity" : ["has_for_level_1_affinity","has_for_level_2_affinity",
    "has_for_level_3_affinity","has_for_level_4_affinity","has_for_level_5_affinity"],

    "abundance" :["has_an_abundance_of_level_1","has_an_abundance_of_level_2",
    "has_an_abundance_of_level_3","has_an_abundance_of_level_4",
    "has_an_abundance_of_level_5"]
  }
}
```

Fig. 6 json file for equality constraints defining equivalence classes of relations on the FRESQUEAU dataset

```
  {
    "Equality" : {
      "constraint1" : ["R1","R2","R3"],
      "constraint2 ": ["R4","R5","R6"]
    }
  }
```

where:
the key `constraint`i, is a string defining the name of a relation set.
the values are the relations with the same scaling quantifier during the process.
$R_{1..6}$ are the relation names;
a given R_i belongs to only one relation set.

Fig. 7 General form of the constraint json file

in Fig. 6 expresses that users want to have the same quantifier on the five abundance relations, the same quantifier on the five PC measures relations and the same quantifier on the five affinity relations. The general form of the *json* file is shown in Fig. 7.

5.2 Interpretation Outline

Another difficulty that we observed during the analysis is understanding the impact of the choice of scaling quantifiers. In order to address this issue, we developed an interpreter which automatically translates the choices made on the user interface into a formatted expression in a controlled language. The box of the upper part of Fig. 8

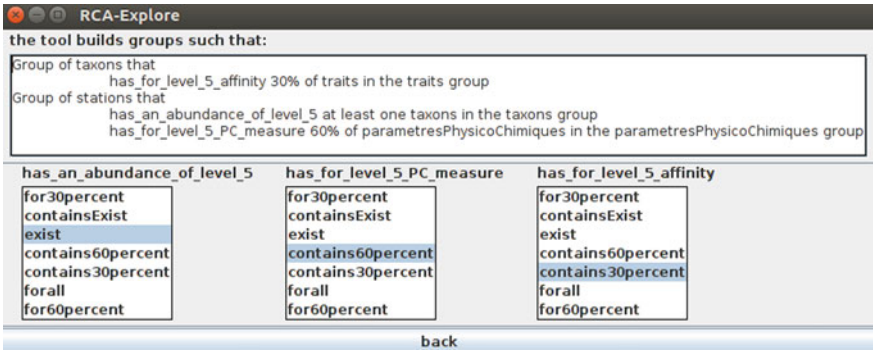

Fig. 8 Scaling quantifier selector and associated interpretation (screenshot excerpt)

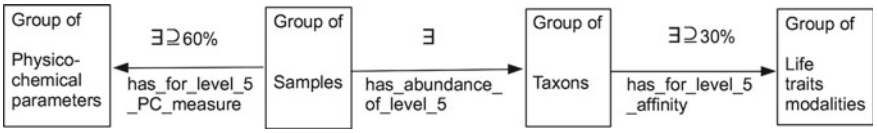

Fig. 9 Schematic representation of quantifier selection for the screenshot excerpt of Fig. 8

shows such an expression corresponding to the selections made in the lower part. It is composed using expressions of the form:

```
Group of < source Formal Context name > that
< Relation name > < quantifier expression >
in the < target Formal Context name > group
```

Such kind of expression corresponds to a group $C = (X, Y) \in \mathscr{L}_{source}$ such that there exists $C' \in \mathscr{L}_{target}$ with $qr.C' \in Y$. C corresponds to `Group of < source Formal Context name >`; r corresponds to `< Relation name >`; q corresponds to `< quantifier expression >`; C' corresponds to `< target Formal Context name > group`.

The `< quantifier expression >` can be one of the following:

`at least one < target Formal Context name >`
`< target Formal Context name > only`
`< n > % of < target Formal Context name >`
`all < target Formal Context name >`
`< n > % of < Relation name > links`

The textual box (upper part of Fig. 8 with an associated schematic interpretation in Fig. 9) is automatically updated when new scaling quantifiers are chosen, allowing experts to immediately capture the meaning of their last choice.

5.3 Dashboard of Neighbor Result Sets

The third overlay consists in computing metrics for neighbor result sets, i.e. potential results of the next step, and was motivated by the difficulty to know the shape (mainly the size) of the answers. The neighborhood is built on the generality relation $\preceq_S$ (more general than) between scaling quantifiers (Braud et al. 2018). It can be based on *forall* ($\exists\forall$ being $\exists\forall_{\geq 100\%}$):

$$\exists\forall \preceq_S (\ldots) \preceq_S \exists\forall_{\geq 60\%}(\ldots) \preceq_S \exists\forall_{\geq 30\%} \preceq_S (\ldots) \preceq_S \exists$$

Or it can be based on *contains* ($\exists\supseteq$ being $\exists\supseteq_{\geq 100\%}$):

$$\exists\supseteq \preceq_S (\ldots) \preceq_S \exists\supseteq_{\geq 60\%}(\ldots) \preceq_S \exists\supseteq_{\geq 30\%} \preceq_S (\ldots) \preceq_S \exists$$

A first metric is defined as the number of concepts that would be computed in each neighbor configuration. Another relevant metric relies on the number of implication rules that can be extracted from the extended formal contexts for each configuration. By implication rule, we mean, in this paper, "attribute implication" in a formal context $K = (G, M, I)$. An implication $X \implies Y$, where $X \subseteq M$ and $Y \subseteq M$ are two attribute sets, holds if the objects that have all the attributes of X also have all the attributes of Y: $\{g \in G | \forall m \in X, (g, m) \in I\} \subseteq \{g \in G | \forall m \in Y, (g, m) \in I\}$.

Several implication rule sets (implicative systems), and more precisely, several bases (non-redundant implicative systems), can be built from a formal context (Bertet et al. 2018). A non-redundant implicative system is a system from which the removal of any of its implicative rules produces a non-equivalent system (a different fact set is deduced). To guide domain experts, we have considered three different implicative systems which highlight different aspects of data.

First, we considered the implication set introduced in Ouzerdine et al. (2019), and we removed the identical ones (they were very few). As these rules are very specific to our dataset, they are introduced in the next section. For new datasets, such specific implication rules should be redefined.

Then, we considered non-redundant implicative systems composed of binary implications, namely, where $|X| = |Y| = 1$. We did not compute these rules, but we evaluated their number. They can be used to recover all binary implications. The cardinal of a non-redundant set of strict binary implications (implications $x \implies y$ such that $y \implies x$ does not hold) can be obtained by counting the number of edges in the transitive reduction of the AC-poset (partially ordered set of concepts introducing at least one attribute). Then, we need to add to the count the cardinal of a non-redundant set of implications $x \implies y$ such that $y \implies x$ also holds. They are given by the simplified intents of the concepts introducing more that one attribute. When a simplified intent contains $n > 2$ attributes, e.g. $\{a, b, c\}$, then an associated non-redundant set of binary implications contains n binary implications, e.g. $a \implies b, b \implies c$, and $c \implies a$. The sum of the cardinals of all simplified intents of attribute introducer concepts introducing more than one attribute gives us the num-

ber of such implications. Besides, the attributes of the bottom concept (if it exists in the AC-poset), when it does not introduce objects, can be removed from the count, as they are not relevant (the support of the corresponding implications is null).

The third non-redundant implicative system that we consider is the best-known basis, namely the *canonical basis*, and has been defined by Guigues and Duquenne and later reworked by several authors (Obiedkov and Duquenne 2007). It is computed thanks to different systems, including Concept Explorer (Conexp)[4] that we used in our experimentation. It is of minimal cardinality, and all the other implication rules can be deduced from the canonical basis using Armstrong axioms (Armstrong 1974).

6 Application to the FRESQUEAU Dataset

In this section, after introducing the part of the FRESQUEAU dataset we used (Sect. 6.1), we describe neighbor result sets for this case: for specific rules in Sect. 6.2, for binary implications in Sect. 6.3, and for the canonical basis in Sect. 6.4.

6.1 FRESQUEAU *Dataset*

The dataset used in this paper, and encoded in a format readable by RCAExplore is available online.[5] It is composed of 1702 water samples collected in Alsace region (East of France), 392 taxons (macroinvertebrates), 116 taxon traits, and 40 physico-chemical parameters. Data were collected during a previous project (Grac et al. 2011) and included in the FRESQUEAU database.

Taxons are described by their name. The initial formal context also indicates taxonomy relationships between taxons (family, gender, etc.). The traits and physico-chemical parameters are also simply described by their name (serving as an identifier). Water samples have no specific attribute (see Fig. 1) thus, their description comes from the relations.

Fifteen relational contexts implement five levels for the relations between objects Water samples and PC measures, between objects Water samples and Taxons, and between objects Taxons and Traits. Levels are based on percentiles for PC measures and abundances, and on predefined levels for affinity, relying on statistical observations (Usseglio-Polatera et al. 2000).

A previous work (Dolques et al. 2016), based on a similar dataset, has focused on decreasing the computational complexity of implication rules (using AOC-posets versus iceberg lattices) and studying the resulting number of concepts and rules with various scaling operators. Only implication rules linking taxon traits and physical characteristics of river streams have been studied. With regard to this work, we

[4] http://conexp.sourceforge.net/.

[5] http://dataqual.engees.unistra.fr/data.

here provide a systematic study of different types of rules, and we explore rules linking physico-chemical measures and taxon abundances in two ways. While work presented in Dolques et al. (2016) focuses on results, here we also describe helping tools for producing and analyzing those results.

6.2 *Specific* FRESQUEAU *Rules*

In our practical study, the aim is to explore the connections between physico-chemical parameters and life traits of taxons. We extract classifications (concept lattices) in which experts navigate, as well as rules. In the metrics presented here, we first compute the number of concepts of the lattices which is a good indicator about the size of the classifications (Table 2, column 3). We also compute the number of particular non-redundant implication rules with premises of size 1 between the relational attributes, that are deduced from introducer concepts and the transitive reduction (Table 2, column 4). In the `Taxon` concept lattice, we chose to extract rules of the following form, to highlight relations between traits:

$$< \texttt{quantified affinity to a life trait group} > \Longrightarrow < \texttt{quantified affinity to a life trait group} >$$

In the `WaterSample` lattice, the extracted rules have the following forms:

$$\texttt{ma:} < \texttt{quantified measure of a PC} - \texttt{parameter} > \Longrightarrow < \texttt{quantified abundance of a taxon group} >$$

$$\texttt{aa:} < \texttt{quantified abundance of a taxon group} > \Longrightarrow < \texttt{quantified abundance of a taxon group} >$$

$$\texttt{mm:} < \texttt{quantified measure of a PC} - \texttt{parameter} > \Longrightarrow < \texttt{quantified measure of a PC} - \texttt{parameter} >$$

$$\texttt{am:} < \texttt{quantified abundance of a taxon group} \Longrightarrow < \texttt{quantified measure of a PC} - \texttt{parameter} >$$

`ma` is supposed to give information on the links between the physico-chemical state of watercourses and the presence of certain groups of taxons. `aa` will reveal the co-presence of taxons. Both rules `aa` and `ma` are consistent with experts' questions. `mm` is supposed to give some already known results, e.g. relations between the various forms of nitrogen, due to chemical processes. `am` is not supposed to give information, since taxons (macroinvertebrates) should not have effect on physico-chemical parameters. The tool could propose to skip the computation of this last type of rule, according to expert knowledge.

The corresponding metrics are shown in Table 2 when tuning n in $\exists\forall_{\geq n\%}$ and $\exists\supseteq_{\geq n\%}$, with $n \in 0, 30, 60, 100$. We recall that $\exists\forall_{\geq 0\%} = \exists\supseteq_{\geq 0\%} = \exists$. In these configurations, we have examined what happened when the scaling quantifier is changed on relation *abundance 3*. The neighbor result sets allow experts to move into the analysis space. Using Table 2 as dashboard, experts first can remark that the dataset contains water samples with a diversity in terms of their PC-composition and the

Table 2 Specific FRESQUEAU rule sets metrics: ∃ affinity and measure, and different operators for abundance

Formal context	Relation	Concept nb	Spec. Fresq. rule nb	Support maximal size
Taxon	∃ affinity 3	460	250	153 on 1 rule
WaterSample	∃ measure 3 ∃ abundance 3	1661	Tot = 415 ma = 19 mm = 3 am = 2 aa = 391	1258 on 1 rule 2 on 19 rules 2 on 3 rules 9 on 1 rule 1258 sur 1 rule
WaterSample	∃ measure 3 $\exists\forall_{\geq 30\%}$ abundance 3	1661	Tot = 433 ma = 32 mm = 3 am = 3 aa = 395	1258 on 1 rule 2 on 32 rules 2 on 3 rules 9 on 1 rule 1258 on 1 rule
WaterSample	∃ measure 3 $\exists\forall_{\geq 60\%}$ abundance 3	1641	Tot = 405 ma = 11 mm = 3 am = 2 aa = 389	1254 on 1 rule 2 on 11 rules 2 on 3 rules 3 on 1 rule 1254 on 1 rule
WaterSample	∃ measure 3 ∃∀ abundance 3	1642	Tot = 408 ma = 7 mm = 3 am = 2 aa = 396	930 on 1 rule 2 on 7 rules 2 on 3 rules 3 on 2 rules 930 on 1 rule
WaterSample	∃ measure 3 $\exists \supseteq_{\geq 30\%}$ abundance 3	1557	Tot = 31 ma = 1 mm = 3 am = 25 aa = 2	2 on 7 rules 2 on 1 rule 2 on 3 rules 1 on 25 rules 1 on 2 rules
WaterSample	∃ measure 3 $\exists \supseteq_{\geq 60\%}$ abundance 3	1514	Tot = 3 ma = 0 mm = 3 am = 0 aa = 0	2 on 3 rules / 2 on 3 rules / /
WaterSample	∃ measure 3 ∃ ⊇ abundance 3	1514	Tot = 3 ma = 0 mm = 3 am = 0 aa = 0	2 on 3 rules / 2 on 3 rules / /

taxons population, explaining the low value of the maximal size of a rule support in many cases. If experts are interested in the concepts and their classification, they can notice that there is not a large difference when they specialize the $\exists\forall$ scaling quantifier (going from 1661 concepts to 1641) but a larger one when they specialize the $\exists\supseteq$ scaling quantifier (going from 1661 concepts to 1514). With the upper part of Table 2, they can notice that lattices for $\exists\forall_{\geq 60\%}$ and $\exists\forall$ have the same number of concepts, making it useless to refine the quantifier up to 100% if their goal is to reduce the size of the result to make it more manageable. If domain experts are more interested in `measure/abundance (ma)` rules, they can consider the number of `ma` rules in column 4, successively 19, 32, 11 and 7. They can decide to use the quantifier $\exists\forall_{\geq 60\%}$ which gives a reasonable number of rules, that are based on a relatively high connection between water samples and taxons. With Table 2 (bottom), they can observe that the quantifiers are too restrictive for the `ma` rules. Besides, it is useless to explore between $\exists\supseteq_{\geq 60\%}$ and $\exists\supseteq$ because the number of rules and the lattice size do not change.

6.3 Non-redundant Binary Implication Sets

Table 3 shows the numbers of non-redundant binary implication sets. All numbers are high, with 1145 (around 10%) more rules when choosing the operator $\exists\forall_{\geq 30\%}$ rather than the operator $\exists$, and 669 (around 5.5%) more when choosing the operator $\exists\forall_{\geq 60\%}$ rather than the operator $\exists\forall_{\geq 30\%}$. We can observe that using $\exists\forall_{\geq 60\%}$ or $\exists\forall$ gives similar results, thus it is useless to explore the operators with intermediate percents on this kind of rules. The rule number is decreased by 500 (around 5%) between $\exists\supseteq_{\geq 30\%}$ and $\exists\supseteq_{\geq 60\%}$, and is little changed afterwards.

6.4 Canonical Basis

Although we have computed the previous indicators without any difficulty on our personal computers, we met difficulties with the canonical basis. We finally computed it on a physical server Dell M630 - 2 x Intel Xeon E5-2697 v3 @ 2.60 GHz - 256 Go RAM and on a virtual server based on the hardware Dell M630 - 2 x Intel Xeon E5-2695 v3 @ 2,30 GHz - 384 Go RAM (using 54 of the 56 available cores, and the whole RAM). Table 4 shows the obtained results after more than 12 h of computation. The other processes were stopped after 170 h without giving any result. Although canonical basis is a relevant information source for experts, this experiment showed that it should not be used in this context.

To conclude this section, whatever metrics are used, our point is that the neighbor result sets allow experts to move into the analysis space, knowing the increase or decrease of the constructed artefacts amount. This will prevent them from using operators that would give no additional information especially when datasets are

Table 3 Binary implication rule sets metrics: ∃ affinity and measure, and different operators for abundance

Formal context	Relation	Concept nb	Binary rule nb
Taxon	∃ affinity 3	460	3580
WaterSample	∃ measure 3 ∃ abundance 3	1661	10982
WaterSample	∃ measure 3 $\exists\forall_{\geq 30\%}$ abundance 3	1661	12127
WaterSample	∃ measure 3 $\exists\forall_{\geq 60\%}$ abundance 3	1641	12796
WaterSample	∃ measure 3 ∃∀ abundance 3	1642	12796
WaterSample	∃ measure 3 $\exists \supseteq_{\geq 30\%}$ abundance 3	1557	10271
WaterSample	∃ measure 3 $\exists \supseteq_{\geq 60\%}$ abundance 3	1514	9771
WaterSample	∃ measure 3 ∃ ⊇ abundance 3	1514	9769

Table 4 Canonical basis metrics: ∃ affinity and measure, and different operators for abundance

Formal context	Relation	Nb of concepts	Nb of rules	Max size support
Taxon	∃ affinity 3	460	38734	291 on 1 rule
WaterSample	∃ measure 3 $\exists\forall_{100\%}$ abundance 3	1642	120872	1262 on 2 rules
WaterSample	∃ measure 3 $\exists \supseteq_{60\%}$ abundance 3	1514	17022	201 on 1 rule
WaterSample	∃ measure 3 $\exists \supseteq_{100\%}$ abundance 3	1514	17022	201 on 1 rule

large or complex, as they may be in the domains we consider. Following this idea, we have recently studied the complexity issue of applying RCA on larger datasets from the hydroecological and agricultural domains (Braud et al. 2020). This study draws the opportunities and the limits of RCA in these cases and concludes with a few new research and technological potential solutions.

7 Conclusion

In this paper, we have presented several overlays that we have integrated in RCA-Explore to guide domain experts without an extended knowledge of RCA in the exploration of datasets. Indeed, RCA allows us to extract classifications, rules and

patterns with a large range of filters (provided by quantifiers), but the modifications that are brought to the results by these filters may not be very intuitive, and the analysis of a large number of generated concepts may be difficult. As a way to limit the choices, the application of constraints allows experts to take into account coherent groups of relations coming from the initial dataset, and to apply quantifiers in a homogeneous way. Besides, the formal writing of the query may be very difficult to understand, possibly leading to wrong choices. The built-in query interpreter to RCAExplore has been implemented, to translate instantly the choices of experts into an advanced language helping the query formulation. The Python scripts used upstream of RCAExplore and applied to the generated lattices, compute metrics that allow experts to have an overview on the extracted information (number of concepts, number and support of extracted rules) and thus to reorient (refine or expand) their search. The generated implication rules inform the experts on the relations between domain objects.

In the future, we plan to develop a version of RCAExplore that will be a fully integrated data exploration tool, allowing to go from raw data to results more easy to exploit by the domain experts. To this aim, an interface should be developed to make transparent the sequencing of the various tool parts, to render all results, some currently being stored in separate files, and to provide a menu assisting the iterative process. Constraint and interpretation languages will be refined to come closer to the natural language, providing a more simple interface. We will also conduct user analyses to observe to which extent this interface and the tool in general facilitates the RCA usage.

Besides, we plan to study the complementarity between RCA and other by-generalization and by-aggregation approaches. We have already achieved comparisons with temporal pattern discovery (Nica et al. 2016) and inductive logic programming (Dolques et al. 2014); other comparisons can be done with pattern structures (Codocedo and Napoli 2014) or bi-clustering, or in general, with machine learning (Kaytoue et al. 2015). These comparisons will be extended to the domain of relational databases, inductive databases, OWL/RDF and SPARQL: Knowledge graphs (such as described with OWL/RDF) can be input data for RCA, as RCA basically considers triplets, and reversely, RCA output (concepts) can be used in ontology design (Hacene et al. 2008). Furthermore, rules discovered by RCA can be included in rule-based systems. Nevertheless the initial purpose of RCA is not reasoning but data mining in the form of knowledge pattern extraction, conceptual classification building and implication/association rule discovery, and more generally machine learning, in the line of FCA (Kuznetsov 2004).

Acknowledgements This research has been partially funded by OFB (Office français de la biodiversité). The authors also thank Corinne Grac (UMR 7362 LIVE - ENGEES) for her valuable help on FRESQUEAU dataset.

References

Armstrong, W. W. (1974). Dependency structures of data base relationships. In *IFIP Congress* (pp. 580–583).

Atencia, M., David, J., Euzenat, J., Napoli, A., & Vizzini, J. (2020). Link key candidate extraction with relational concept analysis. *Discrete Applied Mathematics, 273*, 2–20.

Azmeh, Z., Driss, M., Hamoui, F., Huchard, M., Moha, N., & Tibermacine, C. (2011). Selection of Composable Web Services Driven by User Requirements. In *ICWS* (vol. 2011, pp. 395–402).

Bendaoud, R., Napoli, A., & Toussaint, Y. (2008). Formal concept analysis: A unified framework for building and refining ontologies. In *EKAW 2008. LNCS* (vol. 5268, pp. 156–171).

Bertet, K., Demko, C., Viaud, J., & Guérin, C. (2018). Lattices, closures systems and implication bases: A survey of structural aspects and algorithms. *Theoretical Computer Science, 743*, 93–109.

Bimonte, S., Boulil, K., Braud, A., Bringay, S., Cernesson, F., Dolques, X., Fabrègue, M., Grac, C., Lalande, N., Le Ber, F., & Teisseire, M. (2015). A decisional system for analysing water quality of watercourses. *Revue des Sciences et Technologies de l'Information - Série ISI?: Ingénierie des Systèmes d'Information, 20*(3), 143–167.

Boulil, K., Le Ber, F., Bimonte, S., Grac, C., & Cernesson, F. (2014). Multidimensional modeling and analysis of large and complex watercourse data: An olap-based solution. *Ecological Informatics, 24*, 90–106.

Braud, A., Dolques, X., Gutierrez, A., Huchard, M., Keip, P., Le Ber, F., Martin, P., Nica, C., & Silvie, P. (2020). *Dealing with Large Volumes of Complex Relational Data using RCA*. To be published by Springer.

Braud, A., Dolques, X., Huchard, M., & Le Ber, F. (2018). Generalization effect of quantifiers in a classification based on relational concept analysis. *Knoweldge-Based Systems, 160*(15), 119–135.

Carpineto, C., & Romano, G. (2004). Exploiting the potential of concept lattices for information retrieval with CREDO. *Journal of Universal Computer Science, 10*(8), 985–1013.

Codocedo, V., & Napoli, A. (2014). A proposition for combining pattern structures and relational concept analysis. In C. V. Glodeanu, M. Kaytoue, & C. Sacarea (Eds.), *Formal Concept Analysis* (pp. 96–111). Cham: Springer International Publishing.

Cuenca, E., Sallaberry, A., Ienco, D., & Poncelet, P. (2018). Visual querying of large multilayer graphs. In *Proceedings of the 30th International Conference on Scientific and Statistical Database Management, SSDBM 2018, Bozen-Bolzano, Italy, July 09-11, 2018* (pp. 32:1–32:4).

Dolques, X., Huchard, M., Nebut, C., & Reitz, P. (2012). Fixing generalization defects in UML use case diagrams. *Fundamenta Informaticae, 115*(4), 327–356.

Dolques, X., Le Ber, F., Huchard, M., & Grac, C. (2016). Performance-friendly rule extraction in large water data-sets with AOC posets and relational concept analysis. *International Journal of General Systems, 45*(2), 187–210.

Dolques, X., Le Ber, F., Huchard, M., & Nebut, C. (2015). Relational concept analysis for relational data exploration. *Advances in Knowledge Discovery and Management, 5*, 55–77.

Dolques, X., Mondal, K. C., Braud, A., Huchard, M., & Le Ber, F. (2014). RCA as a data transforming method: a comparison with propositionalisation. In C. V. Glodeanu, C. Sacarea, & M. Kaytoue (Eds.), *ICFCA: International Conference on Formal Concept Analysis*, number 8478 in LNCS (pp. 112–127). Cluj-Napoca: Springer.

Dunaiski, M., Greene, G. J., & Fischer, B. (2017). Exploratory search of academic publication and citation data using interactive tag cloud visualizations. *Scientometrics, 110*(3), 1539–1571.

Džeroski, S. (2003). Multi-relational data mining: An introduction. *ACM SIGKDD Explorations Newsletter, 5*(1), 1–16.

Ferré, S. (2014). *Reconciling Expressivity and Usability in Information Access* (p. 1). Habilitation à diriger des recherches, Université de Rennes.

Ferré, S. (2015). A proposal for extending formal concept analysis to knowledge graphs. In *Formal Concept Analysis, ICFCA 2015, volume LNCS 9113* (pp. 271–286). Spain: Nerja.

Ferré, S., Allard, P., & Ridoux, O. (2012). Cubes of concepts: Multi-dimensional exploration of multi-valued contexts. In *Formal Concept Analysis - 10th International Conference, ICFCA 2012, Leuven, Belgium, May 7-10, 2012. Proceedings* (pp. 112–127).

Ferré, S. & Cellier, P. (2018). How hierarchies of concept graphs can facilitate the interpretation of RCA lattices? In *Proceedings of the Fourteenth International Conference on Concept Lattices and Their Applications, CLA 2018, Olomouc, Czech Republic, June 12-14, 2018* (pp. 69–80).

Ferré, S., Ridoux, O., & Sigonneau, B. (2005). Arbitrary relations in formal concept analysis and logical information systems. In *ICCS'05*, LNAI 3596 (pp. 166–180). Springer.

Ganter, B., & Kuznetsov, S. O. (2001). Pattern structures and their projections. In *ICCS'01, Stanford, CA, USA* (pp. 129–142).

Ganter, B., & Wille, R. (1999). *Formal Concept Analysis: Mathematical Foundations*. Berlin: Springer.

Grac, C., Le Ber, F., Braud, A., Trémolières, M., Bertaux, A., Herrmann, A., Manné, S., & Lafont, M. (2011). Programme de recherche-développement *Indices* – rapport scienfique final. Contrat pluriannuel 1463 de l'Agence de l'Eau Rhin-Meuse, LHYGES – LSIIT – ONEMA – CEMAGREF.

Hacene, M. R., Huchard, M., Napoli, A., & Valtchev, P. (2013). Relational concept analysis: Mining concept lattices from multi-relational data. *Annals of Mathematics and Artificial Intelligence, 67*(1), 81–108.

Hacene, M. R., Napoli, A., Valtchev, P., Toussaint, Y., & Bendaoud, R. (2008). Ontology learning from text using relational concept analysis. In *2008 International MCETECH Conference on e-Technologies (mcetech 2008)* (pp. 154–163).

Ingalalli, V., Ienco, D., & Poncelet, P. (2018). Mining frequent subgraphs in multigraphs. *Information Sciences, 451–452*, 50–66.

Inmon, W. (2005). *Building the Data Warehouse*. Hoboken: Wiley.

Kaytoue, M., Codocedo, V., Buzmakov, A., Baixeries, J., Kuznetsov, S. O., & Napoli, A. (2015). Pattern structures and concept lattices for data mining and knowledge processing. *Machine Learning and Knowledge Discovery in Databases* (pp. 227–231). Berlin: Springer.

Kimball, R., & Ross, M. (2002). *The Data Warehouse Toolkit -The Complete Guide to Dimensional Modeling*. Hoboken: Wiley.

Kötters, J. (2013). Concept Lattices of a Relational Structure. In *ICCS 2013,Mumbai, India*, LNCS, 7735 (pp. 301–310).

Krmelova, M., & Trnecka, M. (2013). Boolean factor analysis of multi-relational data. In *CLA 2013, La Rochelle, France, CEUR Workshop Proceedings* (vol. 1062, pp. 187–198).

Kuznetsov, S. O. (2004). Machine learning and formal concept analysis. In *Concept Lattices, Second International Conference on Formal Concept Analysis, ICFCA 2004, Sydney, Australia, February 23-26, 2004, Proceedings* (pp. 287–312).

Lachiche, N. (2010). Propositionalization. In C. Sammut & G. I. Webb (Eds.), *Encyclopedia of Machine Learning, chapter 17* (pp. 812–817). Berlin: Springer.

Liquière, M., & Sallantin, J. (1998). Structural machine learning with Galois lattice and graphs. In *ICML* (pp. 305–313). Wisconsin: Madison.

Muggleton, S., & Raedt, L. D. (1994). Inductive logic programming: Theory and methods. *Journal of Logic Programming, 19*(20), 629–679.

Nica, C., Braud, A., Dolques, X., Le Ber, F., & Huchard, M. (2016). Exploring temporal data using relational concept analysis – an application to hydroecology. In *Concept lattices and their applications (CLA 2016), CEUR Workshop Proceedings* (pp. 1–13).

Obiedkov, S. A., & Duquenne, V. (2007). Attribute-incremental construction of the canonical implication basis. *Annals of Mathematics and Artificial Intelligence, 49*(1–4), 77–99.

Ouzerdine, A., Braud, A., Dolques, X., Huchard, M., & Le Ber, F. (2019). Régler le processus d'exploration dans l'analyse relationnelle de concepts - le cas de données hydroécologiques. In *Extraction et Gestion des connaissances, EGC 2019, Metz, France, January 21-25, 2019* (pp. 57–68).

Palagi, É., Gandon, F. L., Giboin, A., & Troncy, R. (2017). A survey of definitions and models of exploratory search. In *ACM Workshop ESIDA@IUI* (pp. 3–8).
Prediger, S., & Wille, R. (1999). The lattice of concept graphs of a relationally scaled context. In *ICCS'99, Blacksburg, Virginia*, LNCS 1640 (pp. 401–414). Springer.
Rouane-Hacene, M., Valtchev, P., & Nkambou, R. (2011). Supporting ontology design through large-scale fca-based ontology restructuring. In *ICCS 2011* (pp. 257–269).
Sowa, J. F. (2008). Chapter 5 conceptual graphs. In F. van Harmelen, V. Lifschitz, & B. Porter (Eds.), *Handbook of Knowledge Representation, volume 3 of Foundations of Artificial Intelligence* (pp. 213–237). Elsevier.
Usseglio-Polatera, P., Bournaud, M., Richoux, P., & Tachet, H. (2000). Biomonitoring through biological traits of benthic macroinvertebrates: how to use species trait databases? In M. Jungwirth, S. Muhar, & S. Schmutz (Eds.), *Assessing the Ecological Integrity of Running Waters* (pp. 153–162). Dordrecht: Springer.
Voutsadakis, G. (2002). Polyadic concept analysis. *Order, 19*(3), 295–304.
Wildemuth, B. M., & Freund, L. (2012). Assigning search tasks designed to elicit exploratory search behaviors. In *Human-Computer Information Retrieval Symposium (HCIR)*.
Wolff, K. E. (2009). Relational scaling in relational semantic systems. In *Conceptual Structures: Leveraging Semantic Technologies, 17th International Conference on Conceptual Structures, ICCS 2009, Moscow, Russia, July 26-31, 2009. Proceedings* (pp. 307–320).
Wray, T., & Eklund, P. W. (2011). Exploring the information space of cultural collections using formal concept analysis. In *Formal Concept Analysis - 9th International Conference, ICFCA 2011, Nicosia, Cyprus, May 2-6, 2011. Proceedings* (pp. 251–266).
Zeleny, M. (2013). Integrated knowledge management. *International Journal of Information Systems and Social Change (IJISSC), 4*(4), 62–78.

Amirouche Ouzerdine is a computer scientist since 2007. With an interest in processing health data, he completed in 2018 master studies in bioinformatics. He is currently working on the analysis of metagenomic data with unsupervised analysis methods.

Agnès Braud is assistant professor in computer science at Université de Strasbourg. Her research mainly focuses on data mining from relational and temporal data, using Formal Concept Analysis methods and Inductive Logic Programming.

Xavier Dolques holds a PhD in computer science. He is currently research engineer at ENGEES, Strasbourg, on a research project funded by the French Office of Biodiversity. His research focuses mainly on knowledge discovery from relational data, using Relational Concept Analysis and pattern mining approaches, with applications on environmental data.

Marianne Huchard is professor at LIRMM, Université of Montpellier since 2004. She has been leading research work in FCA for more than 20 years, and contributed to various aspects of this domain: theory, algorithms, methodology and application to several domains, including ontology engineering, environmental datasets, and software engineering.

Florence Le Ber is currently Director of the research department at ENGEES, Université de Strasbourg. She is a senior researcher in computer science, and a specialist of spatial knowledge representation and data mining, including FCA methods. Her research mainly focuses on developing and applying such methods on environmental and agricultural questions.