Dealing with large volumes of complex relational data using RCA

Agnès Braud, Xavier Dolques, Alain Gutierrez, Marianne Huchard, Priscilla Keip, Florence Le Ber, Pierre Martin, Cristina Nica, Pierre Silvie

Chapter 5
Dealing with Large Volumes of Complex Relational Data using RCA

Agnès Braud, Xavier Dolques, Alain Gutierrez, Marianne Huchard, Priscilla Keip, Florence Le Ber, Pierre Martin, Cristina Nica and Pierre Silvie

Abstract Most of available data are inherently relational, with e.g. temporal, spatial, causal or social relations. Besides, many datasets involve complex and voluminous data. Therefore, the exploration of relational data is a major challenge for Formal Concept Analysis (FCA). Relational Concept Analysis (RCA) is specifically designed to investigate the relational structure of a dataset in the FCA paradigm. In this chapter, we examine how RCA can take over the issues raised by complex data. Using two datasets, one about the quality monitoring of waterbodies in France, the other about the use of pesticidal and antimicrobial plants in Africa, we study the limitations of different FCA algorithms, and their current implementations to explore these datasets with RCA. We also show how pattern extraction combined with the presentation of data in hierarchical structures is appropriate for the analysis of temporal datasets by the domain expert. Finally, we discuss about the possible directions to investigate.

Agnès Braud, Xavier Dolques, Florence Le Ber
Université de Strasbourg, CNRS, ENGEES, ICube UMR 7537, F-67000 Strasbourg, France
e-mail: agnes.braud@unistra.fr, [xavier.dolques, florence.leber]@engees.unistra.fr

Alain Gutierrez, Marianne Huchard
LIRMM, Univ Montpellier, CNRS, Montpellier, France
e-mail: [alain.gutierrez, marianne.huchard]@lirmm.fr

Priscilla Keip, Pierre Martin
CIRAD, UPR AIDA, F-34398 Montpellier, France
AIDA, Univ Montpellier, CIRAD, F-34000 Montpellier, France
e-mail: [priscilla.keip, pierre.martin]@cirad.fr

Cristina Nica
Nicolae Titulescu University of Bucharest, Bucharest, Romania
e-mail: nica.cristina87@gmail.com

Pierre Silvie
IRD, UMR IPME, 34AA001 Montpellier, France
CIRAD, UPR AIDA, F-34398 Montpellier, France
AIDA, Univ Montpellier, CIRAD, F-34000 Montpellier, France
e-mail: pierre.silvie@cirad.fr

5.1 Introduction

Many data are inherently relational, and their relations can be complex, numerous, fuzzy and sometimes cyclic. 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). Several approaches have been implemented to explore such data [17]. Relational Concept Analysis (RCA), based on Formal Concept Analysis (FCA), has been specifically designed for this task: it builds a classification (a lattice of formal concepts) for each category of objects contained in a dataset, and allows to obtain implication rules including relations between objects [28, 45].

RCA, as FCA, comes with a major challenge, linked to the fact that dealing with large and complex data produces huge and complex results. Many methods have been proposed to reduce the lattice size, either by reducing the original data (e.g. by granular reduction [55]) or by projection [48], or by reducing the number of concepts to be built (e.g., by thresholding [49]), or by using AOC-posets [15].

Another approach is to help the user to navigate the results, e.g. by focusing on specific subsets of concepts, based on interestingness measures [6, 11], or by using local views and computation on-demand [16, 20]. Regarding RCA, the issue is also to navigate a family of lattices, each concept of a lattice being possibly linked to several concepts of other lattices.

RCA has been applied to multi-relational datasets from various domains, e.g. for the fuzzy semantic annotation of web resources [13], or for the analysis and reengineering of software models [14] and semantic wikis [47]. In previous works, we have applied RCA for exploring hydroecological [15, 38] and agricultural data [29], the two domains considered in the following.

In this paper, we experiment the application of RCA on complex environmental datasets coming from the real world and built under guidance of domain experts. The two application domains are biopesticides and antimicrobial products made from plants (KNOMANA project) and the monitoring of the ecological quality of waterbodies (FRESQUEAU project). In the context of the environmental domains we deal with, the studied datasets can be considered large volume data with regard to the type of data and data collection. In KNOMANA, data are manually collected or revised by experts in scientific publications. The publications are of different types and there is a cross-check in different sources, and a cleaning of information to ensure the data quality. In FRESQUEAU, data are manually collected and manipulated by field biologists in rivers, which differs from data collection from sensors. We show the scope of the RCA process in terms of quantitative opportunities and limits on our datasets, by comparing different algorithms. We also describe an application of RCA to the extraction of graphs from temporal data: the issue is to link sequences of physico-chemical parameter values with bio-indicator values used for assessing the quality of waterbodies. This temporal data pattern extraction shows how we can concretely help domain experts.

As said before, these two datasets have already been studied [15, 29, 38]. In this paper, we propose a synthesis of observations made during these earlier studies,

with enhanced datasets, taking into account more information or applied to the whole initial data rather than to an excerpt. We also compare the algorithms with a same metric set on both datasets.

Section 5.2 presents RCA principles while Sect. 5.3 compares RCA with the related work. Section 5.4 introduces the two complex environmental datasets, and compares the results obtained on these datasets by a few algorithms. Section 5.5 describes the variant of RCA used for analysing sequential datasets. Besides, it shows how summarizing interrelated concepts by a graph can help the analysis of the RCA results. Section 5.6 discusses the results and draws up some perspectives.

5.2 Background

Formal Concept Analysis (FCA) has several dimensions, including being a knowledge engineering method based on lattice theory [25]. In its simplest form, FCA deals with datasets formalized into *formal contexts* comprising objects described by attributes (object-attribute contexts). Attributes in formal contexts are sometimes refered as Boolean attributes. For example, the top of Tab. 5.1 shows four formal contexts: `Biopesticide` describes the toxicity of six biopesticides (from `p1` to `p6`) using two attributes (`toxic`, `nonToxic`); `Bioaggressor` informs on the type of six bioaggressors (from `a1` to `a6`) using two attributes (`worm`, `rodent`); `ProtectedSystem` presents six biological systems to be protected (from `s1` to `s6`) using four attributes (`seed`, `cerealSeed`, `cucurbitSeed`, `leaf`); `Country` localizes four countries (from `c1` to `c4`) in two regions using two attributes (`west`, `east`). A formal context may have a specific shape: it may *partition* the objects with mutually exclusive attributes; it may be *diagonal* if it has the same number of objects and attributes, and each object is described by exactly one attribute (the relation corresponds to a 1-1 mapping).

FCA highlights hierarchies of concepts, each concept being composed of a maximal group of objects (extent) and the maximal group of attributes they share (intent). Since only objects `s3` and `s4` share attributes `seed` and `cerealSeed`, `Concept_ProtectedSystem_2`= (`{s3, s4}`, `{seed, cerealSeed}`) is a concept. For similar reasons, `Concept_ProtectedSystem_4`= (`{s1, s2, s3, s4}`, `{seed}`) is another concept. The set of all concepts provided with inclusion between concept extents (from bottom to top) forms a lattice (the concept lattice). `Concept_ProtectedSystem_2` is a subconcept of `Concept_Protected System_4` in this lattice, as the extent of the former is included in the extent of the latter. Fig. 5.1 shows the concept lattices associated with `Biopesticide`, `Bioaggressor`, `ProtectedSystem`, and `Country`. In this representation of lattices, the attributes (resp. objects) are written only in the highest (resp. lowest) concept where they appear (their introducer concept) and are inherited top to bottom (resp. bottom to top). For instance, `Concept_Biopesticide_2` groups non toxic biopesticides `p1`, `p2`, `p3`, and `p4`, `Concept_Bioaggressor_2` groups worms `a1`,

Biopesticide	toxic	nonToxic
p1		×
p2		×
p3		×
p4		×
p5	×	
p6	×	

Bioaggressor	worm	rodent
a1	×	
a2	×	
a3	×	
a4	×	
a5		×
a6		×

ProtectedSystem	seed	cerealSeed	curcubitSeed	leaf
s1	×		×	
s2	×		×	
s3	×	×		
s4	×	×		
s5				×
s6				×

Country	west	east
c1	×	
c2	×	
c3		×
c4		×

treats	a1	a2	a3	a4	a5	a6
p1						
p2			×			
p3				×		
p4	×	×				
p5					×	
p6						×

attacks	s1	s2	s3	s4	s5	s6
a1	×					
a2		×				
a3			×			
a4				×		
a5					×	
a6						×

isHostedIn	c1	c2	c3	c4
a1			×	
a2				×
a3	×			
a4		×		
a5			×	
a6				×

Table 5.1: Relational Context Family. Top: the formal contexts (object-attribute contexts) `Biopesticide`, `Bioaggressor`, `ProtectedSystem`, `Country`. Bottom: the relational contexts (object-object contexts): `treats`, `attacks`, `isHostedIn`

`a2`, `a3`, and `a4`, `Concept_ProtectedSystem_4` groups seeds `s1`, `s2`, `s3` and `s4`, and `Concept_Country_1` groups western countries `c1` and `c2`.

FCA and all its extensions are well-founded mathematical frameworks thanks to lattice theory, delivering to experts explainable results on which they can base their decisions. FCA is the reference for building exact hierarchies of object/attribute structures, attributes being possibly complex descriptions, and has no competitor for that feature. Uta Priss notes that "the basic FCA structures have been rediscovered over and over by different researchers and in different settings." [46], emphasizing their fundamental aspect. FCA has also a central position as a swiss knife in knowledge engineering and discovery, as the conceptual structures intrinsically contain the search space for rules of different kinds, traceable knowledge patterns and hierarchical structures [43]. FCA is human centered, suitable for interactive and incremental analyses, with visual presentation of extracted patterns. Besides, FCA extensions enable to deal with complex information: numbers, sequences, graphs, temporal data, etc. without converting datasets into simplified formats.

RCA extends the purpose of FCA to relational data, conforming to a conceptual model, e.g. a UML model. We follow up the example with data conforming to the UML model shown in Fig. 5.2: biopesticides treat bioaggressors that attack protected systems; bioaggressors are hosted in countries. This UML model thus structures the dataset into object categories (here biopesticides, bioaggressors, protected systems and countries), objects still being described by attributes. Relationships connect objects of different (or the same) categories: here `treats` connects biopesticides to bioaggressors; `attacks` connects bioaggressors to protected systems; `isHostedIn`

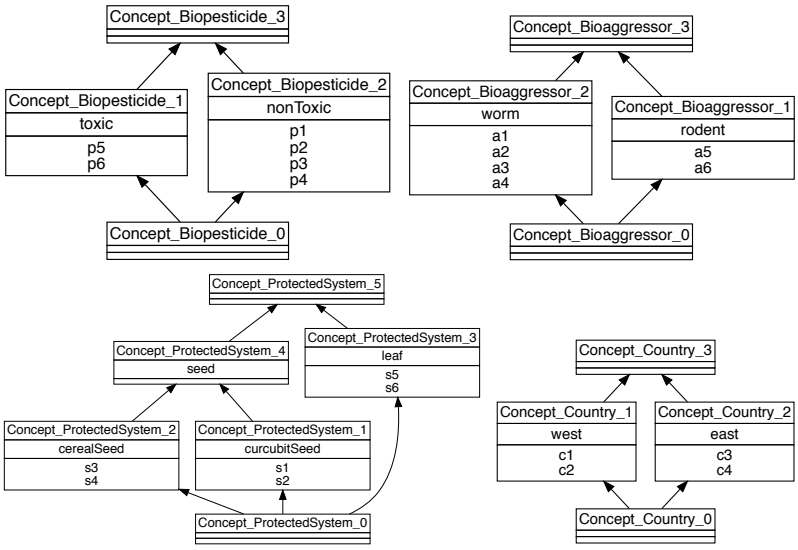


Fig. 5.1: Concept lattices associated with the four formal contexts `Biopesticide`, `Bioaggressor`, `ProtectedSystem`, `Country` from left to right and top to bottom

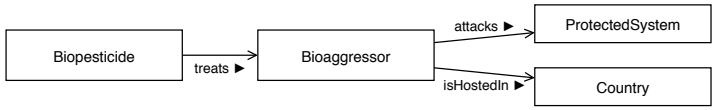


Fig. 5.2: Biopesticides treat bioaggressors that attack protected systems. The bioaggressors are hosted in countries

connects bioaggressors to countries. These relations are shown at the bottom of Tab. 5.1.

The UML model and its instantiation are formalized as a Relational Context Family (RCF). An RCF is a pair $(\mathscr{K},\mathscr{R})$, where $\mathscr{K}$ is a set of object-attribute contexts (formal contexts) and $\mathscr{R}$ is a set of object-object contexts (relational contexts or relations). $\mathscr{K}$ contains n object-attribute contexts $K_i = (G_i, M_i, I_i), i \in \{1,...,n\}$ (*formal contexts*). $\mathscr{R}$ contains m object-object contexts $R_j = (G_k, G_l, r_j)$, $j \in \{1,...,m\}$ (*relational contexts*). $r_j \subseteq G_k \times G_l$ is a binary relation with $k,l \in \{1,...,n\}$. $G_k = dom(r_j)$ is the domain of the relation, and $G_l = ran(r_j)$ is the range of the relation.

The RCA process starts by applying FCA first on each object-attribute context of an RCF. This results in the concept lattices presented in Fig. 5.1.

In the following steps, RCA relies on the construction of particular attributes, called *relational attributes*. These attributes express the relationships an object of one category has with a concept extent (which is a group of objects of a given category). For example, based on `Concept_ProtectedSystem_2` which groups cereal seeds `s3` and `s4`, the relational attribute $\exists attack$(`Concept_Protected-`

System_2), meaning "attack at least one cereal seed", can be formed. This attribute is true for worms a3, a4. This is formalized as follows. A relational attribute $\exists r_j(C)$, where $\exists$ is the existential quantifier, $C = (X,Y)$ is a concept, and $X \subseteq ran(r_j)$, is owned by an object $g \in dom(r_j)$ if $r_j(g) \cap X \neq \emptyset$. Other quantifiers are defined in [9, 28]. In particular, percentage quantifiers are introduced to take into account incomplete, noisy data or approximate satisfaction of a property.

The *relational scaling mechanism* is used to implement the additional description of objects by relational attributes. It maps every relation r_j into a set of *relational attributes* that extend the object-attribute context describing the objects of $dom(r_j)$. This operation is called the relational extension of a context. Table 5.2 shows the relational extension of Bioaggressor at step 1. The first two columns show the original attributes. The next six columns are the relational attributes formed with $\exists$ quantifier, attacks relation, and the concepts of ProtectedSystem lattice of step 0. The next four columns are the relational attributes formed with $\exists$ quantifier, isHostedIn relation, and the concepts of Country lattice of step 0. From this table, worms a3, a4 own relational attributes $\exists attacks$(Concept_ProtectedSystem_2) (cereal seeds) and $\exists isHostedIn$(Concept_Country_1) (western countries).

Bioaggressor	worm	rodent	∃ attacks(Cpt_ProtectedSystem_3)	∃ attacks(Cpt_ProtectedSystem_1)	∃ attacks(Cpt_ProtectedSystem_0)	∃ attacks(Cpt_ProtectedSystem_2)	∃ attacks(Cpt_ProtectedSystem_4)	∃ attacks(Cpt_ProtectedSystem_5)	∃ isHostedIn(Cpt_Country_2)	∃ isHostedIn(Cpt_Country_0)	∃ isHostedIn(Cpt_Country_1)	∃ isHostedIn(Cpt_Country_3)
a1	×			×			×	×	×			×
a2	×			×			×	×	×			×
a3	×					×	×	×			×	×
a4	×					×	×	×			×	×
a5		×	×					×	×			×
a6		×	×					×	×			×

Table 5.2: Relational extension of Bioaggressor at step 1, with relational attributes built on lattices of step 0 (Cpt stands for Concept)

The application of FCA to all the extended contexts refines the original concept lattices. Figure 5.3 shows the Bioaggressor concept lattice at step 1, as a refinement of step 0 (Fig. 5.1). Three concepts are added, in particular Concept_Bioaggressor_4 which groups a1 and a2, that are worms that attack cucurbit seeds and Concept_Bioaggressor_5 which groups a3 and a4, that are worms hosted in western countries and attack cereal seeds. These two concepts emerged thanks to

the addition of relational concepts and divide the group of worms (`Concept_Bioaggressor_2`).

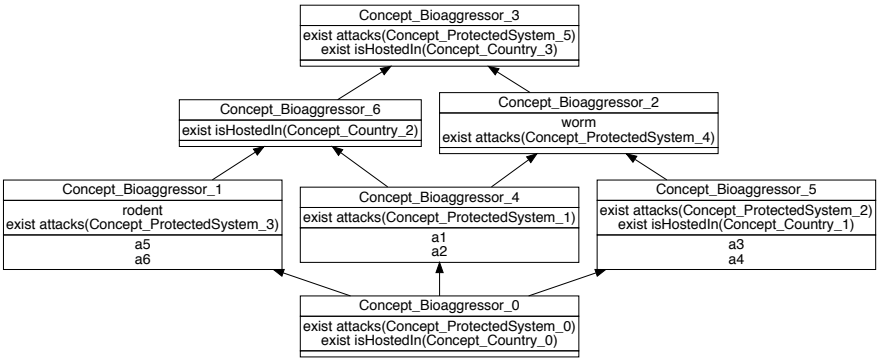


Fig. 5.3: `Bioaggressor` concept lattice at step 1, refining the concept lattice of step 0 (Fig. 5.1)

The complete process operates through successive steps. Each step consists in applying FCA on each object-attribute context extended by the relational attributes created using the concepts from the previous step. This results in a family of concept lattices.

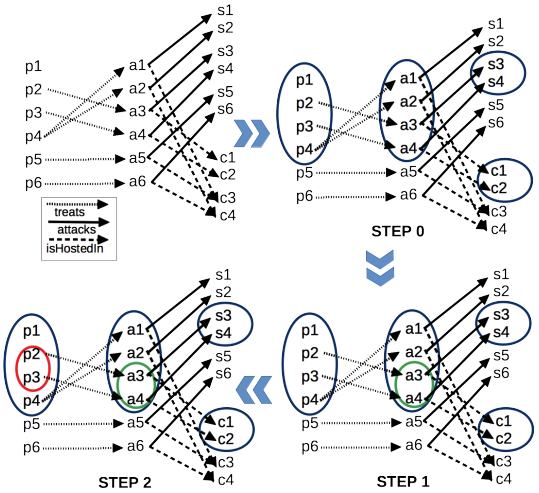


Fig. 5.4: A diagrammatic view on the relational context family with focuses on some concept and sub-concept extents formed at step 0 (top right), step 1 (down right) and step 2 (down left)

Concept formation propagates from one object category to a neighbouring object category along the relations, refining the concept lattices at each step with concept completion, or new concept addition. To continue on our example, Fig. 5.4 represents the relations and objects in the form of a graph (top, left-hand side); it highlights how groups of non toxic biopesticides, worms, cereal seeds and western countries are created at step 0 (blue ellipses). Then, at step 1, one can observe the group of worms from western countries attacking cereal seeds (green ellipse). Then at step 2, the group of non toxic biopesticides allowing to treat them is created (red ellipse). This information appears in the lattice of Fig. 5.5, where `Concept_Biopesticide_7` groups `p2` and `p3`, that are non toxic biopesticides that treat worms hosted in western countries and attacking cereal seeds (through relational attribute $\exists treats$`(Concept_Bioaggressor_5)`).

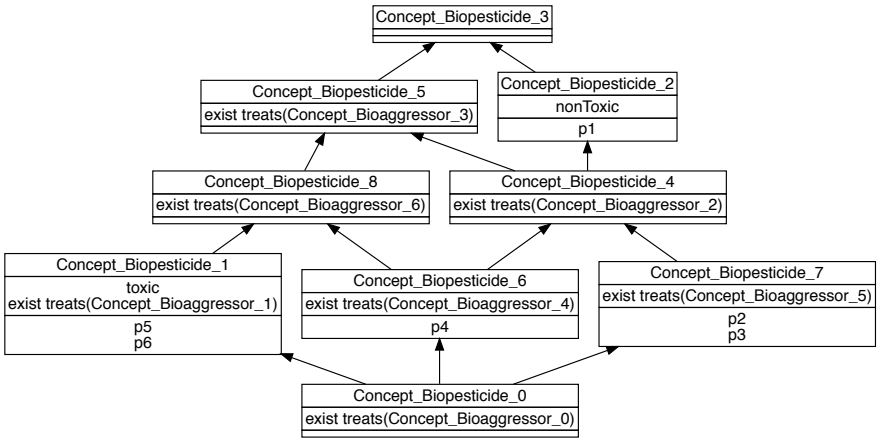


Fig. 5.5: `Biopesticide` concept lattice at step 2 (final step), refining the concept lattice of step 1 (Fig. 5.3)

The RCA process stops when a fixpoint is reached, i.e. when the families of lattices of two consecutive steps are isomorphic and the extended object-attribute contexts are unchanged. The UML model may contain directed cycles, without risk of divergence of the process, as rows (objects) are unchanged, only new columns can be added at each step (with possibly new concepts appearing) and the concept number in each lattice is bounded by $2^{min(|O|,|A|)}$ where O is the object set and A is the attribute set.

5.3 Related Work

Formal Concept Analysis handles multi-relational data through several perspectives. Some approaches extract and classify graph patterns that connect objects or object groups [24,34,44]. Relational data have also been dealt with logical concept analysis [23]. Besides, K. E. Wolff has introduced the 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* [53]. Tuples of Boolean factors are extracted from various tables thanks to an extended version of the Boolean Factor Analysis [31]. An n-ary relation may be in many concrete cases considered as an aggregation of several relations of lower arity. Thus FCA also has been generalized to Triadic Concept Analysis, that considers a ternary relation including objects, attributes and conditions [33]. This yields triadic concepts that are organised in a complete trilattice. This framework has been generalized to n-adic contexts (n-ary relations) in Polyadic Concept Analysis [52].

In [30], a Galois connection (and the derived concept lattice) is introduced to query sets of objects connected by relations. Only existential queries are expressed and there is no iteration. Graph-FCA [21](G-FCA) proposes to consider knowledge graphs based on n-ary relationships as formal contexts. The intent of a G-FCA concept is a projected graph pattern and the extent is an object relation.

Pattern structures [24] can also be used to deal with relational data, for example temporal data. Authors in [10] propose to use pattern structures to build a concept lattice on complex sequential data about care trajectories. The pattern structure is $(P,(S,\sqcap),\delta)$, where P is the set of patients, S is a set of sequences and their sub-sequences, and $\sqcap$ is the set intersection. Each patient of P is described by a sequence (and its sub-sequences) through δ relation. This approach is deepened in [12], where object descriptions are organised into a semi-lattice of closed sets of closed sub-sequences. A similar approach is used for analysing demographic sequences in [26].

Compared to these approaches, RCA benefits from several features. Its derivation from the binary framework makes its results more easy to understand than new diagrams introduced in Relational Semantic Systems. It is relevant for incremental data exploration tasks, as it iterates on knowledge construction, showing the progress in concept construction, contrarily to Boolean factor analysis, pattern structures or Graph-FCA. Compared to the other approaches, it provides several operators to take into account incomplete or noisy data. It has been the subject of research on assisting domain experts in the parametrization and exploration [42].

Several papers push the limits of FCA and show effects of application of FCA or RCA in computation time and conceptual structure size on huge or complex datasets. In [54], a huge Museum collection dataset is analysed and made navigable with FCA, showing the efficacy of the recent algorithms. In [37], RCA is applied to UML class model reengineering, with an underlying circular data schema provoking the construction of large amounts of concepts. Such experiments show that (1) FCA can be applied to huge datasets, (2) RCA, that iterates on FCA, is risky in the presence of cycles and has to be handled with care.

In this paper, we focus on a particular kind of datasets, namely in the environmental domain (with observations, plants, animals, etc.), having in mind that they may present some similarities (in the form of data and the form of querying and exploration needs) and that we should learn some lessons when applying RCA for that specific domain.

5.4 RCA for Environmental Data

In this section, we introduce our datasets, the KNOMANA dataset, and the FRESQUEAU dataset in Sect. 5.4.1. In the following Sect. 5.4.2, we analyse the performances of the current RCA algorithms and implementations on two excerpts of these datasets. After describing the UML model of each excerpt, we give the dimensions of the corresponding context family, the computation time of various RCA algorithms when processing these data, and finally the numbers of concepts and relational attributes of the final lattices. Last section 5.4.3 is a discussion about these results.

5.4.1 Two complex datasets from the environmental domain

5.4.1.1 Pesticidal and Antimicrobial Data

The excessive use of pesticides and antibiotics in agriculture compromises their therapeutic effectiveness and is a threat to human, animal and environmental health [41]. One alternative consists in using natural plant based products. For African farmers, preparing such products using some of the local plants is a challenge. Unfortunately, knowledge on plant use in agriculture is scattered. To support knowledge exchange, description of plants used in Africa was extracted from the scientific literature and collected in a knowledge base called KNOMANA [35]. In KNOMANA, each use of plant is described using 72 data types, among which the protecting plant, the targeted organism (insect, disease, virus, etc.), and the protected system (agricultural crop, animal or human being). In October 2019, KNOMANA gathered 40.800 plant use descriptions for plant, animal, and human health from 410 documents, dated between 1957 and 2019. These uses consider 523 plant protection species, 127 targeted organism species, and 28 protected organism species. In the following (see Sect. 5.4.2) we will explore an excerpt from this database[1].

[1] https://dataverse.cirad.fr/dataverse.xhtml?alias=knomana

5.4.1.2 Water Data

The assessment of aquatic ecosystems, as required by the Water Framework Directive [51], relies on monitoring, which generates large volumes of heterogeneous data from multiple sources at different temporal scales. Actually, when assessing the water quality of watercourses, hydroecologists measure both biological and physico-chemical parameters. In metropolitan France, assessment is done on a network of 1781 sampling sites, called stations. Each station is described by biological data, e.g. the number of individuals for each taxon (animal or plant), and by physico-chemical data, e.g. chemical oxygen demand (denoted DCO), ammonium (denoted NH4), temperature (denoted T), suspended organic matter (denoted MES), etc. Taxons are themselves described by qualitative characteristics, called traits. Based on biological data, biological indicators are computed, e.g. the IBGN ("indice biologique global normalisé") that summarizes information from macro-invertebrate samples into a rating [3], or the IBD ("indice biologique diatomées") that summarizes information from micro-alga samples [2]. Stations are also described by physical and contextual characteristics (e.g. they belong to a waterbody). The assessment varies on time: major physico-chemical parameters are analysed 12 times a year, and minor elements four or six times a year; biological sampling is achieved once a year or once every two years.

Data collected from the 1781 sampling sites from 2007 to 2013 have been recorded into a PostgreSQL/PostGIS database that was designed during the ANR 11 MONU 14 FRESQUEAU project[2]. In the following, we will explore two datasets from this database.

- The first dataset focuses on the annual descriptions of the sampling sites from Jan. 2017 to Nov. 2013: each pair (site, year) is described by the annual average measures of physico-chemical parameters, by taxon lists and by geographical parameters (see Sect. 5.4.2.2).
- The second dataset focuses on the temporal dimension of the data: indeed, each sampling site can be described by a sequence of time stamped physico-chemical parameter measures and time stamped biological indicators (see Sect. 5.5).

All these data are public data, and are freely available on the Naiades (Eau France) website[3].

5.4.2 Experimenting RCA Algorithms

In this section, we assess the ability and limits of RCA and of its current implementation to analyse datasets from the FRESQUEAU and KNOMANA databases.

Both databases can be used in a variety of analyses. To determine the limits of the current RCA implementations, we selected two datasets with representative UML

[2] http://dataqual.engees.unistra.fr/fresqueau_presentation_gb

[3] http://www.naiades.eaufrance.fr/acces-donnees

models. These models were encoded into relational context families. Tables 3 (Sect. 5.4.2.1) and 7 (Sect. 5.4.2.2) describe the formal contexts (object number, attribute number, density) and the relational contexts for each dataset. For the later, only density is indicated as the number of rows and columns results from the source and range formal context object number. The density of a formal context (resp. relational context) is given by the size of the relation (pair number) divided by the object number multiplied by the Boolean attribute number (resp. the source object number multiplied by the range object number).

As the (Boolean) attribute number (column number) corresponds to a scaling of the original (quantitative or multi-valued) attributes, we also indicate these two numbers.

Then the following conceptual structures were built using the $\exists$ quantifier: the concept lattice (addIntent/addExtent algorithm [36]), the AOC-poset (Ceres algorithms [32], Pluton [8], and Hermes [7]), and the Iceberg lattice (Titanic algorithm [50], with minimal support 10%, 30% and 40%). Result tables 4, 5, 6 (Sect. 5.4.2.1) and 8, 9, 10 (Sect. 5.4.2.2) present metrics on the running time, step number, concept number and relational attribute number, and whether computing the structure failed for each dataset. We chose the $\exists$ quantifier, as it generates the largest number of concepts in the worst case, this being the most constraining [9]. Experiments are realized using a laptop with a 4 core Intel i7 2.70 GHz processor.

5.4.2.1 Experiments on KNOMANA Dataset

Figure 5.6 shows the UML model, without cycle, chosen for experimenting RCA algorithms on KNOMANA project. In this model, a `Document` is described by various multi-valued attributes. A document *owns* a piece of `Knowledge` with a certain quality. This piece of knowledge *describes* a form of `HealthProtection` which: *protects* a `ProtectedSystem` *composed of* `ProtectedOrganisms`; *targets* a `TargetedOrganism`; *uses* a `Biopesticide` *made from* a `UsedPlant` from which a technician *extracts* `PlantParts`.

Table 5.3 shows the dimensions of the RCF for the considered excerpt of KNOMANA knowledge base. This RCF is composed of 9 formal contexts and 8 relational contexts. The longest path in the UML model graph is made of 5 edges (from `Document` to `PlantPart`). The largest formal contexts are `HealthProtection` (more than 10000 objects), `Biopesticide` (more than 5000 objects), `UsedPlant` (about 4000 objects), and `Document` (about 3500 objects). Furthermore, Table 3 (and this can also be observed in Table 7 for the FRESQUEAU dataset) shows that the number of objects (rows in formal contexts) is higher than the number of real objects, because an object is described in various situations (e.g. *Lantana camara* may be described in *n* different documents, leading to *n* occurrences of *Lantana camara* (implicitly observations) in the formal context `UsedPlant`). Densities are most of the time low (e.g. `HealthProtection`) to very low (e.g. for *uses*). When the number of objects and attributes are equal, the context may be more complex than a diagonal.

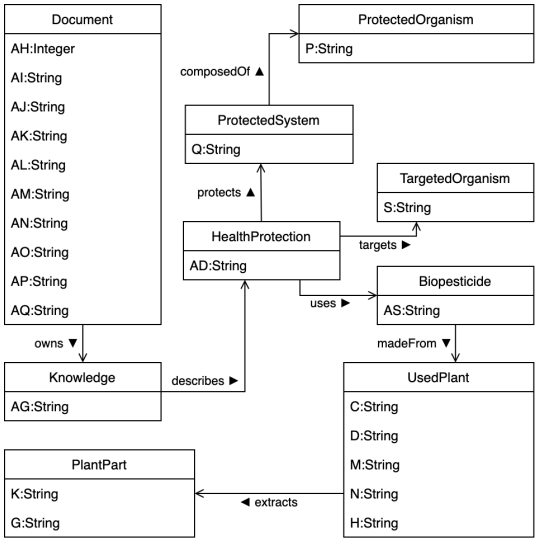


Fig. 5.6: KNOMANA: the UML model used for experimenting RCA algorithms

Formal context	#Objects	#original attributes	#Boolean attributes	density	**Relational context**	density
Document	3541	10	6350	7.73 E-4	**owns**	0.062
Knowledge	16	1	16	0.062	**describes**	0.063
HealthProtection	10172	1	30	0.033	**protects**	0.168
ProtectedSystem	6	1	6	0.167	**composedOf**	0.206
ProtectedOrganism	195	1	195	0.005		
					targets	0.001
TargetedOrganism	934	1	934	0.001		
					uses	3.2 E-4
Biopesticide	3127	1	30	0.033	**madeFrom**	3.23 E-4
UsedPlant	4078	5	4353	7.37 E-4	**extracts**	0.068
PlantPart	344	2	377	0.005		

Table 5.3: KNOMANA: Dimensions of the relational context family

Table 5.4 shows the step number and the running time. The RCA process converged at step 7 (the 7th step being to confirm that the fixpoint is reached). The computing time to construct the concept lattice (FCA) was about 2 minutes. The one for AOC-poset varied between 1 and 8 minutes according to the adopted algorithm, CERES being the most efficient. Iceberg lattices were built in 1 second.

	#steps
FCA	7
CERES	7
PLUTON	7
HERMES	7
ICEBERG10	7
ICEBERG30	7
ICEBERG40	7

	time (ms)	time (mn)
FCA	135582	[Sim] 2
CERES	45355	[Sim] 1
PLUTON	498299	[Sim] 8
HERMES	362232	[Sim] 6
ICEBERG10	517	[Sim] 0.01
ICEBERG30	246	[Sim] 0.01
ICEBERG40	206	[Sim] 0.01

Table 5.4: KNOMANA: (left) Final step number and (right) computation time (milliseconds) and (minutes)

Table 5.5 shows the number of concepts at the final step. The number of concepts in concept lattices varies from one to five times the number of concepts in AOC-posets, except for the `HealthProtection` lattice. In this case, there are 18 times more concepts in the lattice than in the AOC-poset. The concept number of Iceberg lattices is very low compared to the others, suggesting there are no large groups of objects with the same content.

Formal Context	CONCEPT LATTICE	AOC-POSET	ICEBERG10	ICEBERG30	ICEBERG40
BioPesticide	8586	3503	18	11	11
HealthProtection	224992	12182	102	31	15
UsedPlant	5651	4708	8	5	5
ProtectedSystem	23	21	8	8	2
ProtectedOrganism	197	195	2	2	2
PlantPart	386	384	5	2	2
Knowledge	568	116	49	20	8
TargetedOrganism	936	934	2	2	2
Document	6272	4181	74	28	7
TOTAL	247611	26224	268	109	54

Table 5.5: KNOMANA: Number of concepts for each conceptual structure

Table 5.6 shows the number of relational attributes at the final step. For a formal context *Source*, this number is related to the number of concepts of the ranges of the n_{out} relational contexts, i.e. $Range_i,\ 1 \leq i \leq n_{out}$, leaving *Source*. For example, the `Document` relational attributes in the concept lattice (568) originated from the concepts of the `Knowledge` concept lattice. There was a cumulative effect for `HealthProtection`, the relational attributes in the concept lattice (9545) resulting from the union of concepts in `ProtectedSystem` (23), `TargetedOrganism` (936), and `Biopesticide` (8586) concept lattices. This number of relational attributes in `HealthProtection` concept lattice (9545) generated a high number of concepts in `HealthProtection` concept lattice (224992) due to a dispersion of descriptions. During the propagation along *describes* relation, this dispersion was absorbed: 568 concepts only in `Knowledge` concept lattice, and further re-expanded during propagation through *owns* relation in `Document` concept lattice (6272 concepts).

Formal Context	CONCEPT LATTICE	AOC-POSET	ICEBERG10	ICEBERG30	ICEBERG40
BioPesticide	5651	4716	8	5	5
HealthProtection	9545	4458	28	21	15
UsedPlant	386	384	5	2	2
ProtectedSystem	197	195	2	2	2
ProtectedOrganism	0	0	0	0	0
PlantPart	0	0	0	0	0
Knowledge	224992	12464	102	31	15
TargetedOrganism	0	0	0	0	0
Document	568	116	49	20	8
TOTAL	241339	22333	194	81	47

Table 5.6: KNOMANA: Number of relational attributes for each conceptual structure

5.4.2.2 Experiments on FRESQUEAU Dataset

Figure 5.7 shows the UML model, with a loop, chosen for experimenting RCA algorithms on the FRESQUEAU database. As noted above, the dataset represents sampling sites (`Stations`) described by attributes and relations. A station is characterized by a `Year` of observation, it depends on a `WaterBody` and is located in a `HER` (i.e. a Hydro-Eco-Region). A station *is described by* annual average physico-chemical values (`PhCValue`) measured there for 22 parameters (`ParameterName`), and by some fauna or flora lists (`FaunaFloraList`) *containing* the numbers (`TaxonNumber`) of the various *types of* taxons (`Taxon`) collected there at most once a year. Furthermore, taxons have family relationships (`ParentOf`). The aim is to extract groups of stations having similar biological and physico-chemical characteristics.

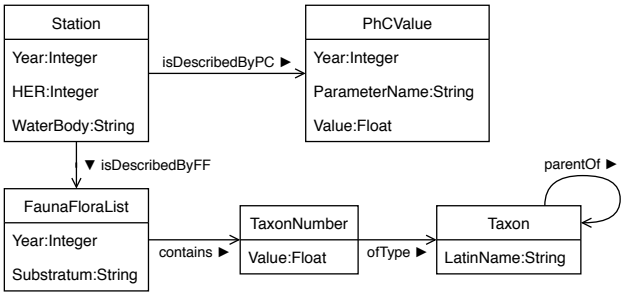


Fig. 5.7: FRESQUEAU: the UML model used for experimenting RCA algorithms

Table 5.7 shows the dimensions of the RCF for the considered excerpt of FRESQUEAU database. This RCF has less contexts but they are larger than those of KNOMANA RCF. Here, the RCF is composed of 4 formal contexts and 5 relational contexts. The longest path (omitting the loop) in the UML model graph has 4 edges (from `Stations` to `Taxon`). The largest contexts are `TaxonNumber` (more than 120000 objects) and `PhCValue` (more than 18000 objects). Original attributes with float values have been discretized into quartiles. Nominal/integer original attributes generated

many Boolean attributes, e.g. 383 for the `Station` context. Taxons are described with no attribute except their name, and thus, `Taxon` context is diagonal. Densities are most of the time low to very low, except for context `TaxonNumber` where there is only 4 Boolean attributes.

Formal context	#Objects	# original attributes	# Boolean attributes	# density	**Relational context**	# density
Station	4808	3	383	0.008	**isDescribedByPC**	0.004
PhCValue	18524	3	512	0.010	**isDescribedByFF**	0.001
FaunaFloraList	510	2	59	0.051	**contains**	1.41E-4
TaxonNumber	124242	1	4	0.25	**ofType**	2.94E-4
Taxon	3400	1	3389	2.95E-4	**parentOf**	2.15E-4

Table 5.7: FRESQUEAU: Dimensions of the relational context family

Table 5.8 shows the step number and the running time for the different algorithms applied to the FRESQUEAU dataset. Concept lattices construction with the add intent/extent algorithm (FCA) cannot finish due to a lack of memory. The RCA process converges in 5 steps for AOC-poset algorithms and Iceberg lattices construction with minimal support 10 (the 5th step being to confirm that the fixpoint is reached). Iceberg lattices construction with minimal supports 30 and 40 needs only 3 steps, due to the fact that less concepts are built and need to be propagated (see Tab. 5.9). AOC-poset construction takes between 4 and 25 minutes depending on the algorithms, CERES being again the most efficient. As for the KNOMANA dataset, Iceberg lattices are easily built in about 1 or 2 seconds.

	#steps
FCA	(-)
CERES	5
PLUTON	5
HERMES	5
ICEBERG10	5
ICEBERG30	3
ICEBERG40	3

	time (ms)	time (mn)
FCA	(-)	(-)
CERES	214577	[Sim] 4
PLUTON	1492604	[Sim] 25
HERMES	979421	[Sim] 16
ICEBERG10	2152	[Sim] 0.04
ICEBERG30	892	[Sim] 0.01
ICEBERG40	711	[Sim] 0.01

Table 5.8: FRESQUEAU: (left) Final step number and (right) computation time (milliseconds) and (minutes)

Table 5.9 shows the number of concepts for each formal context at the final step. The AOC-poset with the highest number of concepts is the one built for `PhCValue`

context. The concept number of Iceberg lattices is low or very low, suggesting, as for the KNOMANA dataset, that there are no large groups of objects with the same description.

Formal Context	CONCEPT LATTICE	AOC-POSET	ICEBERG10	ICEBERG30	ICEBERG40
Station	(-)	1671	89	3	2
PhCValue	(-)	19013	7	2	2
FaunaFloraList	(-)	1171	110	5	5
TaxonNumber	(-)	4524	9	3	3
Taxon	(-)	3392	3	2	2
TOTAL	(-)	29771	218	15	14

Table 5.9: FRESQUEAU: Number of concepts for each conceptual structure

Table 5.10 shows the number of relational attributes for each formal context at the final step. Context `PhCValue` has no relational attribute, not being the domain of any relation. `Station` context is extended with relational attributes pointing to either `PhCValue` or `FaunaFloraList` concepts. Extended contexts `Taxon` and `TaxonNumber` have the same number of relational attributes (the number of concepts of the corresponding `Taxon` lattice, see Tab. 5.9), due to the diagonality of `Taxon` context and the reflexivity of `parentOf` relation.

Formal Context	CONCEPT LATTICE	AOC-POSET	ICEBERG10	ICEBERG30	ICEBERG40
Station	(-)	20321	117	7	7
PhCValue	(-)	0	0	0	0
FaunaFloraList	(-)	4524	9	3	3
TaxonNumber	(-)	3392	3	2	2
Taxon	(-)	3392	3	2	2
TOTAL	(-)	31629	132	14	14

Table 5.10: FRESQUEAU: Number of relational attributes for each conceptual structure

5.4.3 Discussion

In this section, we presented quantitative results about computation time and conceptual structure size for two datasets. We can learn lessons from these experiments on real environmental datasets. Their dimensions present differences, with a very huge object number in one of the formal contexts of FRESQUEAU (`TaxonNumber`). They also have some similarities, with low to very low densities in their formal and relational contexts. The presence of a loop in FRESQUEAU UML model may explain the impossibility to reach the fixpoint for concept lattice construction. Furthermore, when adding all the opposite relations in KNOMANA UML model, some conceptual

structure computation reach the fixpoint: AOC-posets using CERES algorithm are computed within 25 steps in more than one hour and Iceberg 40 lattices are computed within 13 steps in less than a minute. Concept lattices can be built until step 3, Iceberg 10 lattices until step 4 and Iceberg 30 lattices until step 7. This suggests considering other computation strategies to assist domain experts during data exploration tasks: e.g. the process can be interrupted at a certain concept propagation step, which delivers knowledge patterns that can be sufficient for some investigations, or rather than building the whole conceptual structures, concepts can be built around a first focus concept, issued from a set of attributes or objects that have a particular interest for the experts [4].

Furthermore, considering that the conceptual structures are the informative search space for domain experts, raises the question of assisting experts in interpreting and drawing conclusions from the results. This can be made through summarizing patterns (as shown in next section), rule extraction [15,42], or guided exploration of a focus concept neighbourhood [5]. Examples of domain questions for KNOMANA can be found in [29,42]. Next section develops this question with the specific case of the FRESQUEAU project, giving insight on how it can be exploited by an expert to analyse relational data.

5.5 Analysing Sequences from Water Quality Monitoring using RCA

In this section,we consider a smaller but complex relational dataset from FRESQUEAU database (see Sect. 5.4.1.2), that generates large number of concepts when processed by RCA. We show how the lattice family resulting from RCA can be summarized into a single lattice of graphs, to help the interpretation. The approach presented here can be generalized on any relational dataset, possibly larger, thanks to Iceberg lattices construction, and being provided a main lattice to start the summarizing process.

The approach was originally designed to help hydro-ecologists when analysing river water data, and trying to answer the following question: can sets of physico-chemical parameter values be temporally linked with bio-indicator values? To answer this question, Fabrègue et al. [18,19] proposed to use a temporal pattern based method, extracting closed partially ordered patterns (CPO-patterns) from a sequence dataset. Following this idea, and to facilitate the analysis, RCA-Seq has been devised [39,40] for extracting a hierarchy of CPO-patterns from the same datasets. The idea is to represent sequences within a relational context family, to build the lattice family, and then to transform a concept and its related concepts into a CPO-pattern, i.e. a directed acyclic graph (DAG), where each concept corresponds to a vertex, and each relational attribute to an edge. Thus, the lattice family is summarized into a hierarchy of concept graphs [22]. In the following we will explain the functioning of RCA-Seq and present some experiments. Finally we show how the obtained hierarchy can be used to help the expert analysis.

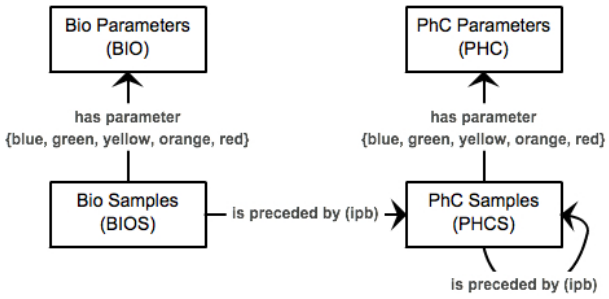


Fig. 5.8: The modelling [38] of hydro-ecological sequential data collected during the FRESQUEAU project; Bio and PhC stand respectively for biological and physico-chemical

5.5.1 RCA-Seq

RCA-Seq spans three main steps: a) modelling sequential data, b) exploring sequential data with RCA and c) extracting DAGs (CPO-patterns) by navigating the RCA output. In the following, we concisely introduce them.

5.5.1.1 Modelling Hydro-Ecological Sequential Data.

In the following a sequence is a list of successive physico-chemical samples collected during a certain period before a biological sampling. To explore these sequential data with RCA, we use the model depicted in Fig. 5.8. The four rectangles represent the four sets of objects we manipulate, as follows: biological (Bio) samples, physico-chemical (PhC) samples, Bio indicators and PhC parameters. Let us note that the analysis only focuses on one Bio indicator at a time. The links between Bio/PhC samples and PhC samples are highlighted by the temporal binary relation *is preceded by*. This temporal relation associates one sample with another one if the first sample is preceded in time by the second one, on the same river stations. Data have been discretised, and the Bio/PhC samples are thus described only by the following binary quality relations *has parameter blue* (very good quality), *has parameter green* (good quality), *has parameter yellow* (medium quality), *has parameter orange* (bad quality) and *has parameter red* (very bad quality) that link the Bio/PhC samples with the measured Bio indicators/PhC parameters.

5.5.1.2 Exploring Hydro-Ecological Sequential Data with RCA.

Firstly, based on the data model given in Fig. 5.8, all sequences, e.g. $\texttt{Seq}_1 = \langle \{\texttt{NITR}_{\texttt{green}}, \texttt{PHOS}_{\texttt{green}}\}\ \{\texttt{NITR}_{\texttt{blue}}\}\ \{\texttt{NITR}_{\texttt{green}}, \texttt{PHOS}_{\texttt{blue}}\}\ \{\texttt{NITR}_{\texttt{green}},$

$\texttt{PHOS}_{\texttt{green}}\}$ $\{\texttt{NITR}_{\texttt{green}}\}$ $\{\texttt{IBGN}_{\texttt{green}}\}\rangle$, are encoded into the RCA input as depicted in Tab. 5.11. The tables KPHC (PhC parameters), KBIOS (Bio samples) and KPHCS (PhC samples) represent object-attribute contexts. There is no object-attribute context for Bio indicators because all analysed sequences end with the same e.g. $\texttt{IBGN}_{\texttt{green}}$. KBIOS and KPHCS have no column since the samples are only described using the quality relations. Moreover they are identified based on the corresponding sequence. The tables RPHCS-ipb-PHCS, RBIOS-ipb-PHCS , RbPHC and RgPHC represent object-object contexts. For example, RPHCS-ipb-PHCS defines the temporal relations (ipb) between PhC samples and has KPHCS both as domain and range. RbPHC (resp. RgPHC) defines the quality relations between PhC samples and PhC parameters that have the blue (b) (resp. the green (g)) quality value.

object-attribute contexts

KPHC	NITR	PHOS
NITR	×	
PHOS		×

KBIOS
$\texttt{Seq}_1$

KPHCS
$\mathfrak{I}_1\_\texttt{Seq}_1$
$\mathfrak{I}_2\_\texttt{Seq}_1$
$\mathfrak{I}_3\_\texttt{Seq}_1$
$\mathfrak{I}_4\_\texttt{Seq}_1$
$\mathfrak{I}_5\_\texttt{Seq}_1$

RBIOS-ipb-PHCS	$\mathfrak{I}_1\_\texttt{Seq}_1$	$\mathfrak{I}_2\_\texttt{Seq}_1$	$\mathfrak{I}_3\_\texttt{Seq}_1$	$\mathfrak{I}_4\_\texttt{Seq}_1$	$\mathfrak{I}_5\_\texttt{Seq}_1$
$\texttt{Seq}_1$	×	×	×	×	×

object-object contexts

RPHCS-ipb-PHCS	$\mathfrak{I}_1\_\texttt{Seq}_1$	$\mathfrak{I}_2\_\texttt{Seq}_1$	$\mathfrak{I}_3\_\texttt{Seq}_1$	$\mathfrak{I}_4\_\texttt{Seq}_1$	$\mathfrak{I}_5\_\texttt{Seq}_1$
$\mathfrak{I}_1\_\texttt{Seq}_1$					
$\mathfrak{I}_2\_\texttt{Seq}_1$	×				
$\mathfrak{I}_3\_\texttt{Seq}_1$	×	×			
$\mathfrak{I}_4\_\texttt{Seq}_1$	×	×	×		
$\mathfrak{I}_5\_\texttt{Seq}_1$	×	×	×	×	

RbPHC	NITR	PHOS
$\mathfrak{I}_1\_\texttt{Seq}_1$		
$\mathfrak{I}_2\_\texttt{Seq}_1$	×	
$\mathfrak{I}_3\_\texttt{Seq}_1$		×
$\mathfrak{I}_4\_\texttt{Seq}_1$		
$\mathfrak{I}_5\_\texttt{Seq}_1$		

RgPHC	NITR	PHOS
$\mathfrak{I}_1\_\texttt{Seq}_1$	×	×
$\mathfrak{I}_2\_\texttt{Seq}_1$		
$\mathfrak{I}_3\_\texttt{Seq}_1$	×	
$\mathfrak{I}_4\_\texttt{Seq}_1$	×	×
$\mathfrak{I}_5\_\texttt{Seq}_1$	×	

Table 5.11: RCF composed of object-attribute contexts: KPHC, KBIOS and KPHCS; temporal object-object contexts: RBIOS-ipb-PHCS and RPHCS-ipb-PHCS; quality object-object contexts: RbPHC and RgPHC; $\mathfrak{I}_k\_\texttt{Seq}_1$ uniquely identifies a PHC sample within a sequence $\texttt{Seq}_1$

Secondly, by applying RCA to an RCF as depicted in Tab. 5.11, three lattices are generated as follows: *the main lattice* $\mathscr{L}_{\texttt{KBIOS}} = (\mathscr{C}_{\texttt{KBIOS}}, \preceq_{\texttt{KBIOS}})$, *the temporal lattice* $\mathscr{L}_{\texttt{KPHCS}} = (\mathscr{C}_{\texttt{KPHCS}}, \preceq_{\texttt{KPHCS}})$ (i.e. PhC samples) and *the lattice of items* $\mathscr{L}_{\texttt{KPHC}} = (\mathscr{C}_{\texttt{KPHC}}, \preceq_{\texttt{KPHC}})$ (i.e. PhC parameters). Throughout the RCA process two types of relational attributes are defined based on the introduced binary relations, either on *time* or *quality*. When $C_m = (X_m, Y_m) \in \mathscr{C}_{\texttt{KBIOS}}$ has $\exists\texttt{RBIOS-ipb-PHCS}(C_{t1}) \in Y_m$, then C_m points to $C_{t1} \in \mathscr{C}_{\texttt{KPHCS}}$. When $C_{t1} = (X_{t1}, Y_{t1})$ has $\exists\texttt{RPHCS-ipb-PHCS}(C_{t2}) \in Y_{t1}$, then C_{t1} points to $C_{t2} \in \mathscr{C}_{\texttt{KPHCS}}$ (i.e. to another concept from the temporal lattice). In addition, when $\exists\texttt{RbPHC}(C_i) \in Y_{t1}$, then C_{t1} points to $C_i = (X_i, Y_i) \in \mathscr{C}_{\texttt{KPHC}}$. When $C_i \equiv \top(\mathscr{L}_{\texttt{KPHC}})$, X_i contains all PhC parameters; for other concepts X_i contains only one PhC parameter.

5.5.1.3 Extracting DAG by Navigating the RCA Output.

As explained in [40], for each concept of the main lattice $C_m = (X_m, Y_m) \in \mathscr{L}_{\texttt{KBIOS}}$ is extracted a DAG (or CPO-pattern), namely $\mathscr{G}_{C_m}$, by navigating interrelated concept intents starting with Y_m; we define $Support(\mathscr{G}_{C_m}) = |X_m|$, i.e. the number of sequences that are summarized by $\mathscr{G}_{C_m}$. All generated DAG are directly organised into a hierarchy according to the inclusion on the associated main concept extents. This order corresponds to the subsumption on graphs $\prec_g$.

Briefly, we explain how to extract a DAG $\mathscr{G}_{C_m} = (\mathscr{V}_m, \mathscr{E}_m, l_m)$ (l_m is a labelling function) associated with a main concept $C_m = (X_m, Y_m) \in \mathscr{C}_{\texttt{KBIOS}}$ whose intent has at least one temporal relational attribute $\exists \texttt{RBIOS-ipb-PHCS}(C_{t1})$, where $C_{t1} = (X_{t1}, Y_{t1}) \in \mathscr{C}_{\texttt{KPHCS}}$. Concept C_m reveals a vertex $v_m \in \mathscr{V}_m$ labelled with an itemset containing the assessed Bio indicator, e.g $\{\texttt{IBGN}_{\texttt{green}}\}$. The aforementioned temporal relational attribute leads to another vertex $v_{t1} \in \mathscr{V}_m$ derived from C_{t1}, i.e. the edge $(v_{t1}, v_m) \in \mathscr{E}_m$ is disclosed. If a quality relational attribute $\exists \texttt{RbPHC}(C_i) \in Y_{t1}$ with $C_i = (X_i, Y_i) \in \mathscr{C}_{\texttt{KPHC}}$, then v_{t1} is labelled with Y_i. Precisely, if $C_i \equiv \top(\mathscr{L}_{\texttt{KPHC}})$, then the *abstract quality value* $?_{\texttt{blue}} \in l(v_{t1})$ is derived; if $C_i \prec_{\texttt{KPHC}} \top(\mathscr{L}_{\texttt{KPHC}})$ with e.g. $Y_i = \{\texttt{PHOS}\}$, then the *concrete quality value* $\texttt{PHOS}_{\texttt{blue}} \in l(v_{t1})$; if Y_{t1} has no quality relational attribute, then the *abstract value* $?_? \in l(v_{t1})$. If Y_{t1} contains a temporal relational attribute $\exists \texttt{RPHCS-ipb-PHCS}(C_{t2})$, then C_{t1} leads to another vertex $v_{t2} \in \mathscr{V}_m$ derived from C_{t2}. Therefore, the order on vertices in $\mathscr{G}_{C_m}$ is revealed by temporal relational attributes; the itemsets labelling the vertices are revealed by quality relational attributes. When all next navigated concept intents have no temporal relational attribute, then the extraction of $\mathscr{G}_{C_m}$ is finished.

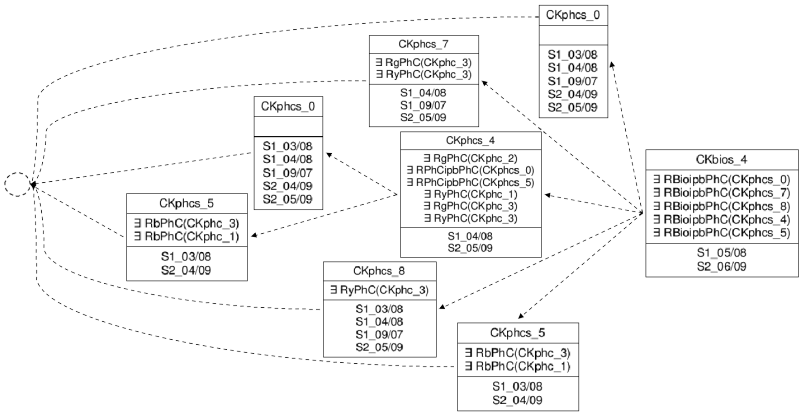


Fig. 5.9: Extracting a DAG by navigating the relational attributes starting from a KBIOS concept; the objects in the extents represent (station, time stamp) pairs, the intents of $C_{\texttt{KPHC\_N}}$ concepts are parameters, NITR (N=1), PHOS (2), any (3)

Figure 5.9 illustrates the extraction of a DAG starting from a concept $C_{\texttt{KBIOS\_4}}$ (right of the figure). This concept has 5 temporal relational attributes that lead to 5 concepts of the lattice $\mathscr{L}_{\texttt{KPHCS}}$. Some of these concepts have quality relational attributes (e.g. $\exists \texttt{RgPHC}(C_{\texttt{KPHC\_3}})$ for concept $C_{\texttt{KPHCS\_7}}$) leading to concepts of $\mathscr{L}_{\texttt{KPHC}}$; while others (e.g. $C_{\texttt{KPHCS\_4}}$) have temporal relational attributes that lead again to concepts of $\mathscr{L}_{\texttt{KPHCS}}$ (left of the figure). Since these last concepts ($C_{\texttt{KPHCS\_5}}$, $C_{\texttt{KPHCS\_0}}$) have no temporal relational attributes, the extraction is finished.

5.5.2 Experiments

This section presents an experimental study of our approach. The experiments were carried out on a MacBook Pro with a 2.9 GHz Intel Core i7, 8GB DDR3 RAM running OS X 10.9.5. The family lattice was built with RCAExplore[4]. The extraction step relied on the CPOHrchy algorithm from [40].

To assess the performance of RCA-Seq we used two hydro-ecological sequential datasets, IBD blue and IBGN blue, whose characteristics, i.e, number of sequences, number of PHC samples, number of PHC parameters, average sequence length (the number of PHC samples in the sequence), maximum sequence length and density, are shown in Tab. 5.12. Figures 5.10(a) and 5.10(d) depict the number of obtained DAGs (vertical axis) with respect to the minimum support θ (%) (horizontal axis) in the IBD and IBGN blue dataset. Even if both datasets have almost the same number of sequences, the extracted number of DAGs varies. For instance, 300411 DAGs are discovered in the IBGN blue dataset with $\theta = 9\%$, while for the same minimum support in the IBD blue dataset only 16525 DAGs are discovered. This difference can be linked to each dataset heterogeneity.

Dataset	#sequences	#PHC samples	#PHC parameters	Avg. seq. length	Max. seq. length	Density
IBD blue	1196	3012	46	2.51	7	2.37E-4
IBGN blue	1102	3077	26	2.79	8	3.17E-4

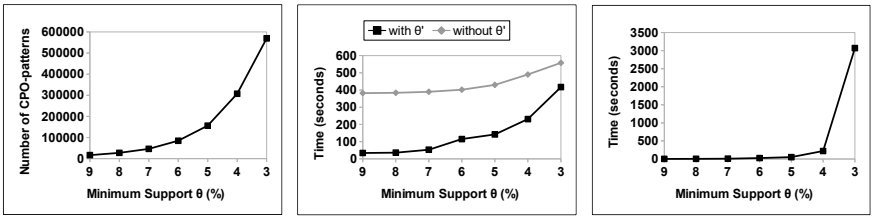
Table 5.12: Dataset characteristics

The number of extracted DAGs is important, even if the dataset is rather small, e.g. we report a number of 569202 DAGs discovered with $\theta = 3\%$ for the IBD dataset that contains only 1196 sequences built from 46 items and having an average sequence length of 2.51 (Fig. 5.10(a)).

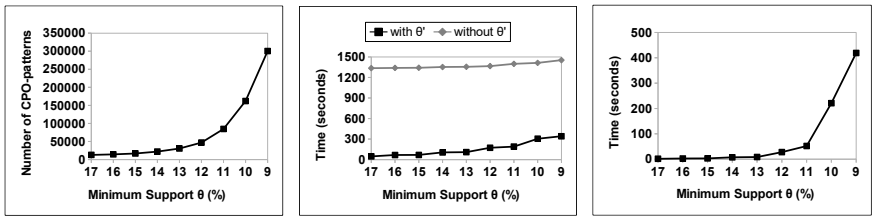
Figure 5.10(b) illustrates the execution time of the RCA-based exploration. As explained in [40], to optimise RCA-Seq we defined respectively for the lattices $\mathscr{L}_{\texttt{KBIOS}}$ and $\mathscr{L}_{\texttt{KPHCS}}$ the minimum supports θ and θ', where $\theta' = \theta \frac{|G_{\texttt{KBIOS}}|}{|G_{\texttt{KPHCS}}|}$. $G_{\texttt{KBIOS}}$ and $G_{\texttt{KPHCS}}$ are respectively the set of objects of the KBIOS and KPHCS object-attribute contexts. For instance, when θ' is not defined (i.e. non-optimised RCA-exploration), during the iterative steps the relational scaling mechanism processes $|\mathscr{C}_{\texttt{KPHCS}}| = 105850$ temporal concepts even if not all of them are used to extract DAGs. When $\theta = 6\%$ and $\theta' = 3\%$, only $|\mathscr{C}_{\texttt{KPHCS}}| = 4429$ temporal concepts are generated; when $\theta = 3\%$ and $\theta' = 1\%$, then $|\mathscr{C}_{\texttt{KPHCS}}| = 31854$ temporal concepts are generated. Thus, for $\theta = 6\%$ and $\theta = 3\%$ the optimised RCA-based exploration is respectively 3.49 and 1.33 times faster than the non-optimised one.

Figure 5.10(c) shows the computation time of the algorithm CPOHrchy. It is noted that low values of θ, i.e. $< 4\%$, and high numbers of DAGs, i.e. ≥ 300000, slow down the extraction step. In addition, the efficiency of CPOHrchy can be influenced

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



(a) DAGs (IBD) (b) RCA-exploration (c) CPOHrchy



(d) DAGs (IBGN) (e) RCA-exploration (f) CPOHrchy

Fig. 5.10: Performance evaluation based on IBD and IBGN blue datasets of Tab. 5.12; minimum support θ is defined for $\mathscr{L}_{\texttt{KBIOS}}$; θ' is defined for $\mathscr{L}_{\texttt{KPHCS}}$

by the used implementation[5], which is not currently optimised for searching in large collections.

5.5.3 Navigating the Resulting Hierarchy of Graphs

Figure 5.11 depicts an excerpt from a hierarchy of DAGs extracted by applying RCA-Seq to an IBGN blue dataset with 80 analysed hydro-ecological sequences. This excerpt highlights two benefits of exploring qualitative sequential data by means of RCA. Firstly, the generalisation order regarding the structure of the extracted DAGs. For example, the structure of DAG (e) is more specific than the structure of its ancestor DAGs, i.e. there exists a projection from its ancestor DAGs into (e). Secondly, the partial order on items, and the inclusion order on itemsets. For instance, DAG (g) reveals the regularity $\{\texttt{MOOX}_{\texttt{blue}}\} \leftarrow \{\texttt{IBGN}_{\texttt{blue}}\}$ (i.e. a blue quality of IBGN Bio indicator is frequently preceded by a blue quality of MOOX PhC parameter) that is a specialisation of the less accurate regularity $\{?_{\texttt{blue}}\} \leftarrow \{\texttt{IBGN}_{\texttt{blue}}\}$ (i.e. a blue quality of IBGN Bio indicator is frequently preceded by a blue quality of some PhC parameter) revealed by DAG (b). Similarly, DAG (i) reveals $\{\texttt{PHOS}_{\texttt{blue}}, \texttt{NITR}_{\texttt{blue}}\} \leftarrow \{\texttt{IBGN}_{\texttt{blue}}\}$ regularity that is a specialisation of the regularity $\{\texttt{PHOS}_{\texttt{blue}}\} \leftarrow \{\texttt{IBGN}_{\texttt{blue}}\}$ revealed by (e). In addition, DAG (e), having 12.5% frequency (i.e. in Fig. 5.11, DAG

[5] based on Java Collection Framework and Lambda Expressions

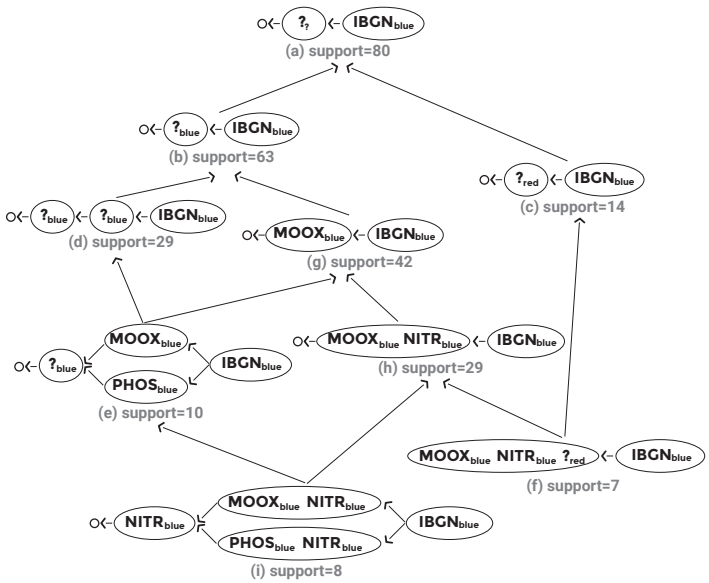


Fig. 5.11: Excerpt from a hierarchy of DAGs generated from an IBGN blue dataset with 80 analysed hydro-ecological sequences

(e) has $Support = 10$, and the total number of analysed sequences is equal to 80), can be found when, e.g. $\theta = 12\%$, even if its accurate specialization DAG (i) is not frequent, and thus is not extracted. These properties of the extracted hierarchies help experts in understanding the obtained knowledge, and, besides, provide a quick way to navigate to interesting DAGs.

Accordingly, the hierarchy in Fig. 5.11 can be navigated starting from the more general DAGs. Thus, the experts have an overview of the trends within analysed data, and minimize the chance of overlooking interesting ones. DAG (a) confirms that all analysed Bio samples are preceded by at least one PhC sample. Both direct descendants, the (b) and (c) DAGs, emphasize two well-known correspondence between the qualities of PhC parameters and the ones of IBGN Bio indicator. Firstly, DAG (b), which is retrieved with 78.75% frequency in the analysed data, highlights that $\texttt{IBGN}_{\texttt{blue}}$ is frequently preceded by PhC parameters having *blue* qualities. Secondly, DAG (c), which is retrieved in 17.5% of the analysed data, stresses that *red* PhC parameters are not frequently measured before an $\texttt{IBGN}_{\texttt{blue}}$ since they produce a degradation of the watercourse qualities, and do not lead to a very good ecological status. As expected, in contrast with DAG (b), DAG (c) has a low support. Therefore, the experts can navigate only descendants of DAG (b) in order to find patterns revealing pertinent synergies between PhC parameters and $\texttt{IBGN}_{\texttt{blue}}$ that provide a very good quality of watercourses. In addition, the experts can focus only on descendants of (c) to find out how the watercourse degradation is neutralised when *red* PhC parameters are measured. For example, the neutralisation of $?_{\texttt{red}}$ is possible when

these very bad values of PhC parameters coexist at the same time with the good values $\texttt{MOOX}_{\texttt{blue}}$ and $\texttt{NITR}_{\texttt{blue}}$ as shown by DAG (f). Following the same principles, the experts can continue the navigation being guided by the relationships between the extracted DAGs and the information about their support.

5.6 Conclusion

In this paper, we have presented results on the application of RCA to environmental datasets coming from the real world and built under guidance of domain experts. The two application domains are biopesticides and antimicrobial products made from plants and assessment of the quality of waterbodies. We have shown the scope of the RCA process in terms of quantitative opportunities and limits on our datasets. We also have described qualitative results in the second domain, about the challenging issue of temporally linking physico-chemical parameter values with bio-indicator values.

We are pursuing two related main tracks of research: (1) improving time and space efficiency of the RCA implementation and adding new algorithmic strategies; (2) improving guidance of experts in their analysis.

Concerning track (1), finding opportunities for *space and time efficiency improvement* is a main and complex task, with tangled concerns, both theoretical and technical. We are experimenting various collection types as many libraries exist that can have an impact on efficiency: Java API collections (currently `BitSet` is used as a main provider for efficiency in experiments of Sect. 5.4), colt library [6] (used in experiments of Sect. 5.5), Apache common collection library [7], Google Guava [8], etc. Other construction algorithms for the concept lattice will be implemented as well, such as described and experimented in [1, 54]. Another technical but important concern is about the data, whose input and output file format (currently textual input format, and dot output format, with optionnally XML output file unfeasible on large results) and memory encoding (currently adjacency lists) have an impact on efficiency. We also are designing on-demand and local algorithms for RCA, following the first work presented in [5]. There are plenty of different ways to consider an on-demand local algorithm, and this way of computing and delivering results has a strong potential for complex, large and evolving datasets.

Concerning track (2), *guidance of experts* can be strenghten by various means. RCA quantifiers offer many possibilities for analysis, with the counterpart that the expert may be lost when choosing parameters (quantifiers and conceptual structures). To that aim, we are studying assisting methodologies, by providing a controlled language for expressing a general "query" with different quantifiers associated with the various relations, and the possible choice of only parts of the RCF; by controlling the

[6] https://dst.lbl.gov/ACSSoftware/colt/

[7] https://commons.apache.org/proper/commons-collections/

[8] https://github.com/google/guava/wiki

coherence between the quantifier choices on semantically connected relations; and by anticipating the result size on neighbouring configurations (with slight changes in the analysed RCF part, or with "similar" quantifiers). The other challenge is providing a user interface with result visualization adapted to the domain experts. Presenting concept orders is used in many tools, while others focus on a particular concept and allow navigating to its neighbours [20], or give an alternative view on the conceptual structure through tag clouds [27].

Acknowledgments. This work was supported by the French National Research Agency: (1) FRESQUEAU project referred as ANR11_MONU14; (2) KNOMANA project under the Investments for the Future Program, #Digitag, referred as ANR-16-CONV-0004. KNOMANA project is also supported by INRA-CIRAD GloFoodS metaprogram. FRESQUEAU database was completed thanks to the support of French Office of Biodiversity (OFB). We warmly acknowledge Corinne Grac (UMR7362 LIVE - ENGEES) for her advice about the FRESQUEAU datasets.

References

1. Andrews, S.: Making use of empty intersections to improve the performance of cbo-type algorithms. In: Formal Concept Analysis - 14th International Conference, ICFCA 2017, Rennes, France, June 13-16, 2017, Proceedings, pp. 56–71 (2017). DOI 10.1007/978-3-319-59271-8_4
2. Association Française de Normalisation: Qualité de l'eau : détermination de l'Indice Biologique Diatomées (IBD). NF T90-354 (2003)
3. Association Française de Normalisation: Qualité de l'eau : détermination de l'Indice Biologique Global Normalisé (IBGN). XP T90-350 (2004)
4. Bazin, A., Carbonnel, J., Huchard, M., Kahn, G.: On-demand relational concept analysis. CoRR **abs/1803.07847** (2018). URL http://arxiv.org/abs/1803.07847
5. Bazin, A., Carbonnel, J., Huchard, M., Kahn, G., Keip, P., Ouzerdine, A.: On-demand relational concept analysis. In: Formal Concept Analysis - 15th International Conference, ICFCA 2019, Frankfurt, Germany, June 25-28, 2019, Proceedings, pp. 155–172 (2019). DOI 10.1007/978-3-030-21462-3_11
6. Belohlavek, R., Macko, J.: Selecting important concepts using weights. In: P. Valtchev, R. Jäschke (eds.) Formal Concept Analysis: 9th International Conference, ICFCA 2011, Nicosia, Cyprus, May 2-6, 2011. Proceedings, pp. 65–80. Springer Berlin Heidelberg (2011)
7. Berry, A., Gutierrez, A., Huchard, M., Napoli, A., Sigayret, A.: Hermes: a simple and efficient algorithm for building the AOC-poset of a binary relation. Ann. Math. Artif. Intell. **72**(1-2), 45–71 (2014). DOI 10.1007/s10472-014-9418-6
8. Berry, A., Huchard, M., McConnell, R.M., Sigayret, A., Spinrad, J.P.: Efficiently computing a linear extension of the sub-hierarchy of a concept lattice. In: B. Ganter, R. Godin (eds.) Formal Concept Analysis, Third International Conference, ICFCA 2005, Lens, France, February 14-18, 2005, Proceedings, *Lecture Notes in Computer Science*, vol. 3403, pp. 208–222. Springer (2005). DOI 10.1007/978-3-540-32262-7_14
9. Braud, A., Dolques, X., Huchard, M., Le Ber, F.: Generalization effect of quantifiers in a classification based on relational concept analysis. Knowledge-Based Systems **160**, 119–135 (2018)
10. Buzmakov, A., Egho, E., Jay, N., Kuznetsov, S.O., Napoli, A., Raïssi, C.: On Mining Complex Sequential Data by means of FCA and Pattern Structures. Int. Journal of General Systems **45**, 135–159 (2016)

11. Buzmakov, A., Kuznetsov, S.O., Napoli, A.: Is Concept Stability a Measure for Pattern Selection? Procedia Computer Science **31**, 918–927 (2014)
12. Codocedo, V., Bosc, G., Kaytoue, M., Boulicaut, J.F., Napoli, A.: A proposition for sequence mining using pattern structures. In: Proceedings of the 14th Int. Conf. on Formal Concept Analysis, ICFCA, pp. 106–121. Springer (2017)
13. De Maio, C., Fenza, G., Gallo, M., Loia, V., Senatore, S.: Formal and relational concept analysis for fuzzy-based automatic semantic annotation. Applied Intelligence **40**(1), 154–177 (2014)
14. Dolques, X., Huchard, M., Nebut, C., Reitz, P.: Fixing Generalization Defects in UML Use Case Diagrams. Fundam. Inform. **115**(4), 327–356 (2012)
15. 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). DOI 10.1080/03081079.2015.1072927
16. Ducrou, J., Eklund, P.: SearchSleuth: The Conceptual Neighbourhood of an Web Query. In: CEUR Workshop Proceedings: Concept Lattices and their Applications, vol. 331, pp. 249–259 (2007)
17. Džeroski, S.: Multi-relational data mining: an introduction. ACM SIGKDD Explorations Newsletter **5**(1), 1–16 (2003)
18. Fabrègue, M., Braud, A., Bringay, S., Grac, C., Le Ber, F., Levet, D., Teisseire, M.: Discriminant temporal patterns for linking physico-chemistry and biology in hydro-ecosystem assessment. Ecological Informatics **24**, 210–221 (2014). DOI 10.1016/j.ecoinf.2014.09.003. URL https://hal.archives-ouvertes.fr/hal-01090331
19. Fabrègue, M., Braud, A., Bringay, S., Le Ber, F., Teisseire, M.: Mining closed partially ordered patterns, a new optimized algorithm. Knowledge-Based Systems **79**, 68 – 79 (2015). DOI https://doi.org/10.1016/j.knosys.2014.12.027. URL http://www.sciencedirect.com/science/article/pii/S0950705114004730
20. Ferré, S.: Camelis: a logical information system to organise and browse a collection of documents. Int. J. General Systems **38**(4), 379–403 (2009). DOI 10.1080/03081070902857886
21. Ferré, S.: A Proposal for Extending Formal Concept Analysis to Knowledge Graphs. In: 13th Int. Conference, ICFCA 2015, Nerja, Spain, LNCS 9113, pp. 271–286 (2015)
22. Ferré, S., Cellier, P.: How hierarchies of concept graphs can facilitate the interpretation of RCA lattices? In: 14th Int. Conference CLA 2018, Olomouc, Czech Republic, pp. 69–80 (2018)
23. Ferré, S., Ridoux, O., Sigonneau, B.: Arbitrary Relations in Formal Concept Analysis and Logical Information Systems. In: 13th Int. Conf. on Conceptual Structures, ICCS'05, Kassel, Germany, LNAI 3596, pp. 166–180. Springer (2005)
24. Ganter, B., Kuznetsov, S.O.: Pattern structures and their projections. In: 9th Int. Conference ICCS'01, Stanford, CA, USA, pp. 129–142 (2001)
25. Ganter, B., Wille, R.: Formal concept analysis - mathematical foundations. Springer (1999)
26. Gizdatullin, D., Ignatov, D.I., Mitrafanova, E., Muratova, A.: Classification of demographic sequence basd on pattern structures and emerging patterns. In: Supplementary Proceedings of ICFCA, pp. 49–66 (2017)
27. Greene, G.J., Esterhuizen, M., Fischer, B.: Visualizing and exploring software version control repositories using interactive tag clouds over formal concept lattices. Information & Software Technology **87**, 223–241 (2017). DOI 10.1016/j.infsof.2016.12.001
28. 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)
29. Keip, P., Gutierrez, A., Huchard, M., Le Ber, F., Sarter, S., Silvie, P., Martin, P.: Effects of input data formalisation in relational concept analysis for a data model with a ternary relation. In: Formal Concept Analysis - 15th International Conference, ICFCA 2019, Frankfurt, Germany, June 25-28, 2019, Proceedings, pp. 191–207 (2019). DOI 10.1007/978-3-030-21462-3_13
30. Kötters, J.: Concept Lattices of a Relational Structure. In: 20th Int. Conf. ICCS 2013, Mumbai, India, LNCS 7735, pp. 301–310 (2013)
31. Krmelova, M., Trnecka, M.: Boolean Factor Analysis of Multi-Relational Data. In: CLA 2013, La Rochelle, France, CEUR Workshop Proc. 1062, pp. 187–198 (2013)

32. Leblanc, H.: Sous-hiérarchie de Galois : un modèle pour la construction et l'évolution des hiérarchies d'objets. Ph.D. thesis, Université de Montpellier (2000)
33. Lehmann, F., Wille, R.: A triadic approach to formal concept analysis. In: 3rd Int. Conference ICCS'95, Santa Cruz, California, USA, pp. 32–43 (1995)
34. Liquière, M., Sallantin, J.: Structural Machine Learning with Galois Lattice and Graphs. In: ICML, Madison, Wisconsin, pp. 305–313 (1998)
35. Martin, P., Sarter, S., Tagne, A., Ilboudo, Z., Marnotte, P., Silvie, P.: Knowing the useful plants for organic agriculture according to literature: Building and exploring a knowledge base for plant and animal health. In: African organic conference, pp. 137–141 (2018)
36. van der Merwe, D., Obiedkov, S.A., Kourie, D.G.: Addintent: A new incremental algorithm for constructing concept lattices. In: 2nd Int. Conference ICFCA 2004, Sydney, Australia, pp. 372–385 (2004)
37. Miralles, A., Molla, G., Huchard, M., Nebut, C., Deruelle, L., Derras, M.: Class model normalization - outperforming formal concept analysis approaches with aoc-posets. In: Proceedings of the Twelfth International Conference on Concept Lattices and Their Applications, Clermont-Ferrand, France, October 13-16, 2015, pp. 111–122 (2015). URL http://ceur-ws.org/Vol-1466/paper09.pdf
38. Nica, C., Braud, A., Dolques, X., Huchard, M., Le Ber, F.: Exploring Temporal Data Using Relational Concept Analysis: An Application to Hydroecology. In: M. Huchard, S. Kuznetsov (eds.) CLA: Concept Lattices and their Applications, *CEUR Workshop Proceedings*, vol. 1624, pp. 299–311. Moscow, Russia (2016). URL https://hal.archives-ouvertes.fr/hal-01380404
39. Nica, C., Braud, A., Dolques, X., Huchard, M., Le Ber, F.: Extracting Hierarchies of Closed Partially-Ordered Patterns Using Relational Concept Analysis. In: 22nd Int. Conf. ICCS 2016, Annecy, France, pp. 17–30 (2016)
40. Nica, C., Braud, A., Le Ber, F.: RCA-Seq: an Original Approach for Enhancing the Analysis of Sequential Data Based on Hierarchies of Multilevel Closed Partially-Ordered Patterns. Discrete Applied Mathematics **273**, 232–251 (2020). DOI 10.1016/j.dam.2019.02.037
41. O'Neill, J.: Tackling drug-resistant infections globally: final report and recommendations. Review on Antimicrobial Resistance. Wellcome Trust and the Department of Health of United Kingdom (2016). 80 pages
42. Ouzerdine, A., Braud, A., Dolques, X., Huchard, M., Le Ber, F.: Adjusting the exploration flow in Relational Concept Analysis. An experience on a watercourse quality dataset. To appear in Advances in Knowledge Discovery and Management, Springer
43. Poelmans, J., Kuznetsov, S.O., Ignatov, D.I., Dedene, G.: Formal Concept Analysis in knowledge processing: A survey on models and techniques. Expert Syst. Appl. **40**(16), 6601–6623 (2013). DOI 10.1016/j.eswa.2013.05.007
44. Prediger, S., Wille, R.: The Lattice of Concept Graphs of a Relationally Scaled Context. In: 7th Int. Conf. on Conceptual Structures, ICCS'99, Blacksburg, Virginia, LNCS 1640, pp. 401–414. Springer (1999)
45. Priss, U.: Relational concept analysis: Semantic structures in dictionaries and lexical databases. Ph.D. thesis, Technische Universität Darmstadt (1996)
46. Priss, U.: Formal concept analysis in information science. ARIST **40**(1), 521–543 (2006). DOI 10.1002/aris.1440400120
47. Shi, L., Toussaint, Y., Napoli, A., Blansché, A.: Mining for Reengineering: An Application to Semantic Wikis Using Formal and Relational Concept Analysis, pp. 421–435. Springer Berlin Heidelberg, Berlin, Heidelberg (2011)
48. Singh, P.K.: m-polar fuzzy graph representation of concept lattice. Engineering Applications of Artificial Intelligence **67**(Supplement C), 52–62 (2018)
49. Stumme, G.: Efficient data mining based on formal concept analysis. In: Database and Expert Systems Applications, pp. 534–546. Springer (2002)
50. Stumme, G., Taouil, R., Bastide, Y., Pasquier, N., Lakhal, L.: Computing iceberg concept lattices with Titanic. Data & Knowledge Engineering **42**(2), 189–222 (2002)
51. The European Parliament and the Council: Framework for Community action in the field of water policy. Directive 2000/60/EC (2000)

52. Voutsadakis, G.: Polyadic concept analysis. Order **19**(3), 295–304 (2002). DOI 10.1023/A:1021252203599
53. Wolff, K.E.: 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 (2009). DOI 10.1007/978-3-642-03079-6_24
54. Wray, T., Outrata, J., Eklund, P.W.: Scalable Performance of FCbO Algorithm on Museum Data. In: Proceedings of the Thirteenth International Conference on Concept Lattices and Their Applications, Moscow, Russia, July 18-22, 2016, pp. 363–376 (2016). URL http://ceur-ws.org/Vol-1624/paper28.pdf
55. Wu, W.Z., Leung, Y., Mi, J.S.: Granular computing and knowledge reduction in formal contexts. IEEE Transactions on Knowledge and Data Engineering **21**(10), 1461–1474 (2009)