Relational Concept Analysis for Relational Data Exploration

Xavier Dolques, Florence Le Ber, Marianne Huchard, Clémentine Nebut

Relational Concept Analysis for Relational Data Exploration

Xavier Dolques, Florence Le Ber, Marianne Huchard, Clémentine Nebut

Abstract Relational Concept Analysis (RCA) is an extension to the Formal Concept Analysis (FCA) which is an unsupervised classification method producing concept lattices. In addition RCA considers relations between objects from different contexts that allow for the creation of links between lattices. This feature makes it more intuitive to extract knowledge from relational data and gives richer results. However, data with many relations imply scalability problems and results that are difficult to exploit. We propose in this article a possible adaptation of RCA to explore relations in a supervised way in order to increase the performance and the pertinence of the results.

1 Introduction

Formal Concept Analysis [Ganter and Wille(1999)], written shortly FCA, is an automatic classification method of objects described by attribute through a binary relation. Such a classification results in a concept lattice (also called Galois lattice [Barbut and Monjardet(1970)]) where each concept groups all the objects sharing the same attribute set. It is possible to navigate through a lattice in a simple and intuitive way, from the most specific concepts (concepts grouping many character-

Xavier Dolques,
ICUBE, Université de Strasbourg/ENGEES, CNRS, Strasbourg, e-mail: xavier.dolques@engees.unistra.fr
Florence Le Ber,
ICUBE, Université de Strasbourg/ENGEES, CNRS, Strasbourg, e-mail: florence.leber@engees.unistra.fr
Marianne Huchard,
LIRMM, Université de Montpellier 2 et CNRS, Montpellier, e-mail: marianne.huchard@lirmm.fr
Cl'ementine Nebut,
LIRMM, Université de Montpellier 2 et CNRS, Montpellier, e-mail: clementine.nebut@lirmm.fr

istics shared by only a few objects) to the less specific ones (concepts grouping many objects but sharing only a few characteristics).

FCA is used in several domains as a knowledge extraction method and the different publications on the topic, namely [Carpineto and Romano(2004), Valtchev et al(2004)Valtchev, Missaoui, and Godin] have identified its forces and limitations. Some of those limitations can be worked around by using different approaches.

Relational Concept Analysis (RCA) [Huchard et al(2007)Huchard, Hacène, Roume, and Valtchev] is an extension of FCA taking into account relations between objects in addition to the characteristics of the objects. RCA consists in iteratively applying an FCA algorithm to deal with relational data: objects are described by attributes and by their relations towards other objects. Concepts discovered by a given iteration are propagated along the relations, leading to the discovery of new concepts at the next iteration.

RCA appears to be more intuitive to use on relational data such as databases or object-oriented modeling languages such as UML. In this article we propose to adapt RCA for the purpose of using it as a knowledge extraction method on water quality measures data for Alsacian watercourses.

This work is part of the ANR project FRESQUEAU [1] which goal is to develop new study, comparison and exploitation approaches of all the available parameters on watercourses. It extends a previous study using FCA [Bertaux et al(2009a)Bertaux, Le Ber, Braud, and Trémolières] and statistical approaches [Bertaux et al(2009b)Bertaux, Le Ber, Li, and Trémolières].

Propagating along relations the concepts discovered from an iteration to another permits to discover interesting concepts, but it often leads to a combinatorial explosion of the number of concepts, and the interesting patterns are difficult to extract from the big concept set built. Several strategies can be used to counter this complexity, including the separation of the initial objects in several subsets after preliminary analysis or the introduction of requests [Azmeh et al(2011)Azmeh, Huchard, Napoli, Hacene, and Valtchev]. We are interested in this article in using RCA to explore interactively the data by letting the user choose before each iteration of FCA which contexts (object-attribute and object-object) he or she wants to use.

We are working on data that are not initially shaped as a binary relation but many works about data scaling will permit to get a binary relation [Ganter and Wille(1999)] or pattern structures [Ganter and Kuznetsov(2001)]. Those approaches have been previously applied on similar data in [Bertaux et al(2009a)Bertaux, Le Ber, Braud, and Trémolières] therefore in the following we only consider data as binary relations.

In this paper we present FCA, then the general principle of the RCA process as to highlight several variation points that would permit to improve the use of RCA in a data mining context. We then present an example of the kind of data from the FRESQUEAU project and the consequences of the variations on these data. We then conclude with a short discussion.

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

2 Formal Concept Analysis

FCA's purpose as defined by [Ganter and Wille(1999)] is to classify a set of objects described by attributes and presented as a formal context. A formal context $\mathscr{K}$ is a triplet (O,A,I) where O is an object set, A is an attribute set and $I \subseteq O \times A$ is the incident relation between O and A such that $(o,a) \in I$ if and only if a is an attribute of o. Table 1 represent a formal context. The object set is here a set of identifiers for sampling stations on different watercourses. Each station is represented by a row. The attributes are description characteristics of the watercourses. The relation between a station and a characteristic of its watercourse is represented by a cross. Thus the station identified by `BREI0001` is located in a `small watercourse` which water is `fresh and live`. The stations `BRUN001` and `BRUN002` are located on the same river but at different locations.

	small watercourse	large watercourse	fresh and calm water	fresh and live water	phreatic watercourse
BREI0001	x			x	
BRUMB001	x		x		
BRUN001					x
BRUN002					x
DOLL001	x			x	
FECH001		x		x	

Table 1 Example of formal context. Objects are presented as rows and attributes as columns.

Applying FCA on a context $\mathscr{K} = (O,A,I)$ leads to the generation of concepts. A concept is a couple (X,Y) where $X \subseteq O$ and $Y \subseteq A$ such that $X = \{o \in O | \forall a \in Y, (o,a) \in I\}$ and $Y = \{a \in A | \forall o \in X, (o,a) \in I\}$. X is called the *extent* of the concept Y its *intent*. The extent of a concept is the maximal set of objects sharing the intent attributes and the intent of a concept is the maximal set of attributes shared by all the extent objects.

For a given context FCA leads to the generation of all the concepts. Those concepts are forming a concept lattice also called Galois lattice. A concept c_1 is more general (resp. more specific) than a concept c_2 if the extent of c_1 contains (resp. is contained by) the extent of c_2. In a dual way, the intent of a concept is contained by the intent of a concept more specific. Two given concepts have a unique superior bound and a unique inferior bound.

Lattices are usually represented by their Hasse diagram. The lattice of table 1 is represented by figure 1. Arrows are representing the generalization relation, *i.e.* the pointed concept is more general than the concept of the origin. Considering that the intent of a concept is included by the intent of every concept more specific and the extent of a concept is included by the extent of every concept more general, each object (resp. attribute) is displayed only once in the most specific (resp. most general) concept where it appears. For instance, concept 6 groups stations `BREI0001` and

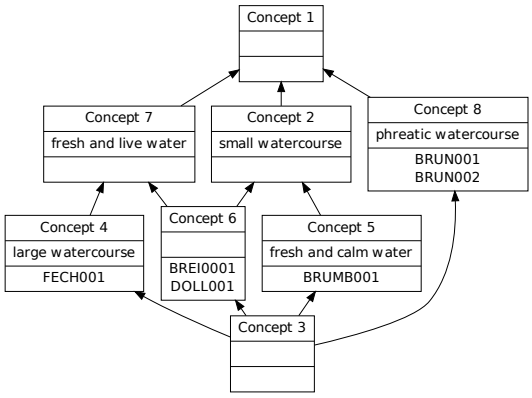

Fig. 1 Hasse diagram of the concept lattice generated from the context described by table 1.

`DOLL001` that own attributes `small watercourse` and `live and fresh water` that can be found by the generalization relation towards concepts 2 and 7.

3 RCA Extension for exploratory analysis

Relational Concept Analysis (RCA) [Huchard et al(2007)Huchard, Hacène, Roume, and Valtchev] is an extension of FCA considering, in addition to object characteristics, existing relations between objects.

The algorithm 1 present the main steps of RCA. The input parameter for RCA is a Relational Context Family $RCF = (K,R)$ composed of n object-attribute contexts $\mathscr{K}_i = (O_i,A_i,I_i)$, $i \in [1..n]$, and m object-object contexts $\mathscr{R}_j = (O_k,O_l,I_j)$, $j \in [1..m]$ where O_k and O_l are object sets of $\mathscr{K}_k$ et $\mathscr{K}_l$. It can be seen in table 2 an example of a Relational Context Family. It can be seen on the left hand side two object-attribute contexts `taxons` and `stations` and on the right hand side the object-object context `taxonPresence` that links objects from context `stations` to objects of context `taxons`[2].

For $\mathscr{R}_j \subseteq O_k \times O_l$, we call O_k its domain and O_l its range. The initialization step (lines 4-5) consists in building, for all $i \in [1..n]$, the lattice $\mathbf{L}^0[i]$ associated to the context $\mathscr{K}_i$. Figure 2 present the two lattices obtained after the initialization step on our example. It can be noticed that the relation `taxonsPresence` is not considered at this point of the processus and that the two lattices are independent.

At step p:

- **EXTEND-REL** add to $\mathscr{K}_i$ the relations obtained by scaling the relations where $\mathscr{K}_i$ is the domain. The scaling consists in the inclusion of the object-object rela-

[2] The term *taxon* covers diverse terms used for the denomination of living beings such as species, genus or families.

```
1: proc MULTI-FCA( Input: (K,R) a RCF,
2: Out: L table [1..n] of lattices)
3: p ← 0 ; halt ← false
4: for i from 1 to n do
5:   L^0[i] ← BUILD-LATTICE(K_i^0)
6: while not halt do
7:   p++
8:   for i from 1 to n do
9:     K_i^p ← EXTEND-REL(K_i^{p-1}, L^{p-1})
10:    L^p[i] ← UPDATE-LATTICES(K_i^p, L^{p-1}[i])
11:  arrłt ← ⋀_{i=1,n} ISOMORPHIC(L^p[i], L^{p-1}[i])
```

Algorithme 1: *Processus of Relational Concept Analysis.*

object-attribute contexts | object-object contexts

taxons	≤1 year	> 1 year
Athericidae	x	
Bithynia	x	x
Boreobdella		x

stations	small watercourse	fresh and live watercourse	phreatic watercourse
BREI0001	x	x	
BRUN001			x
FECH001		x	

taxonPresence	Atheri--cidae	Bithy--nia	Boreob--della
BREI0001			x
BRUN001	x	x	
FECH001	x		x

Table 2 Relation Context Family example. Objects are presented as rows and attributes as columns.

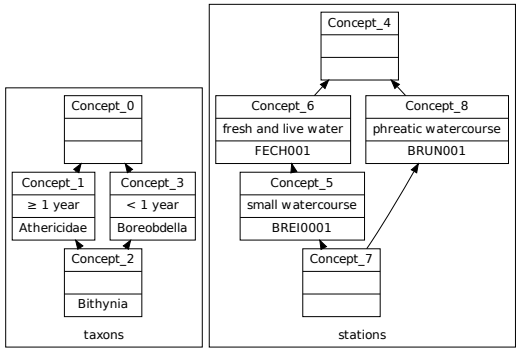

Fig. 2 Lattice Family generated from the Relational Context Family of table 2 after initialization step 0.

stations	small watercourse	fresh and live watercourse	phreatic watercourse	∃*taxonPresence* Concept_0	Concept_1	Concept_2	Concept_3
BREI0001	x	x		x			x
BRUN001			x	x	x	x	x
FECH001		x		x	x		x

Table 3 Scaling of the relation *taxonPresence* and extension of context *stations* at step 1.

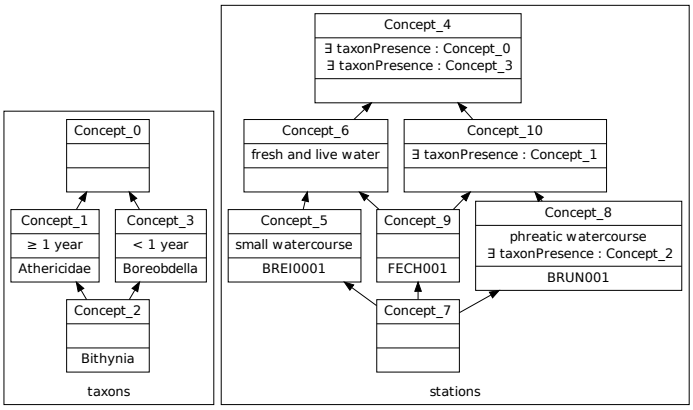

Fig. 3 Lattice Family generated from the Relational Context Family of table 2 after step 1.

tions as *relational attributes*. They are obtained by using lattices concepts from step $p-1$ and a scaling operator (i.e. $\exists$, $\forall$). For example, if the scaling operator $\exists$ is chosen to scale a given relation $\mathscr{R}_j$, the columns of $\mathscr{R}_j$ are replaced by attributes of the form $\exists\mathscr{R}_j : C$, where C is a concept in the lattice built from objects of the range of $\mathscr{R}_j$ at step $p-1$. An object o from the domain of $\mathscr{R}_j$ owns the relational attribute $(\exists\mathscr{R}_j : C)$ if $\mathscr{R}_j(o) \cap Extension(C) \neq \emptyset$. Thus, we have extended the object-attribute context `stations` with the object-object context `taxonsPresence` scaled by the operator $\exists$. The extended context is presented by the table 3. The station `FECH001` is linked to concepts 0, 1 and 3 by the relation $\exists$*taxonsPresence* as we can found the presence of `Athericidae` that can be found in concepts 0 and 1 and the presence of `Boreobdella` that can be found in concepts 0 and 3. If we had used the scaling operator $\forall$, the station `FECH001` would be linked only to concept 0 by the relation $\forall$*taxonsPresence* as it is the only concept where we can find both taxons for this station.

- **UPDATE-LATTICE** update the lattices of step $p-1$ to generate, for $i \in [1..n]$, the lattice $\mathbf{L}^p[i]$, associated to $\mathscr{K}_i$ concatenated to every scaled object-object context which domain is $\mathscr{K}_i$.

The algorithm stops when a fix point is reached, i.e. when the lattice family obtained is isomorphic to the one from the previous step and the context extensions is unchanged. In our example, the lattices from figure 3 are the final lattices. The number of iterations is predictable when relations between contexts are not forming

a circuit. But in some cases, for instance when an object-object context has same domain and range, the number of iterations is not predictable (only a maximal bound can be known) and can be really big depending on the data.

The relational lattices interpretation is different from the interpretation of classical concept lattices as they must be considered silmutaneously. The lattice `stations` of figure 3 must be considered with the lattice `taxons` to be correctly interpreted. We find in concept intents some attributes referring to other concepts. E.g. `Concept_8` owns the relational attribute `taxonsPresence : Concept_2` which means that all the objects of `Concept_8` are linked by the relation `taxonsPresence` with at least (as the scaling operator used is $\exists$) an object of `Concept_2` extent from lattice `taxons`.

The advantage of such a process is that the concepts obtained have in their intent relations to other concepts in addition to classical attributes. Those relations permit the extraction of patterns built from several interconnected contexts, as it has previously been done in [Dolques et al(2009)Dolques, Huchard, and Nebut] and [Dolques et al(2010)Dolques, Huchard, Nebut, and Reitz], that could not be easily obtained from the classical FCA process.

However a major drawback of this kind of process is the potential difficulty to apprehend the result. In previous works in Model Driven Engineering, the data extracted from models of medium size can easily be apprehended by RCA. However, in a data mining context, the data size is more important. Computing time depends on the number of generated concepts, and it is exponential in regard to the minimum between the number of attributes or objects in the worst cas. Thus, if the relations between objects are numerous and with few similarities between objects, computation time can exponentially increase and the result may appear difficult to understand by a user because of the number of concepts to consider simultaneously. This is particularly true when only small patterns are needed when a lot of relations link the objects and these relations are forming a circuit. In such cases, we think it would be relevant to use and exploratory approach.

We list in the following the different possible variations on the algorithm to put in practice an exploratory approach. We enumerate the possible variation points in the algorithm that can affect the result by changing the contexts considered at each step. We propose for each variation point an alternate scenario from the previously described process that involve the user by asking him to choose. All those variations or only a subset of them can be applied depending on the needed granularity.

- **initialization step, line 4 to 5** Build lattices for selected object-attribute contexts concatened to selected object-object contexts.
- **EXTEND-REL, line 9** Instead of using all the relations and scaling all the object-object relation at each step, select a subset of the Relational Context Family and different scaling operators for each object-object context selected. Notice : the object-object relations need a lattice classifying the object of the range that must have been computed in a previous step, not necessarily $p-1$. At this step, object-attribute context can also be selected and the corresponding lattice can be built.
- **UPDATE-LATTICES, line 10** Update only the lattices for the selected relations.
- **stop, line 11** If a fix point is not reached, let the stop decision to the expert.

4 Exploration example

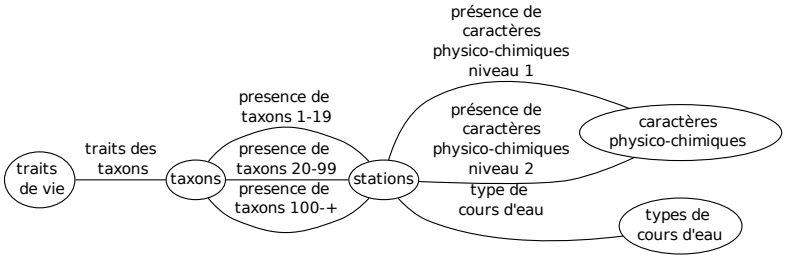

Fig. 4 Schema of the data used by our example.

In this section we illustrate the potential of an exploratory approach using the example of alsacian watercourses. Each watercourse is classified depending on general types (e.g. un small watercourse with fresh and live water). To evaluate the water quality of a watercourse, hydro-ecologists select several sections, called stations, on which they take samples and measures in addition to the sampling of the present plants and animals (called taxons). Those samples and measures respect several norms. After analysis or determination in a laboratory, the watercourse stations are described by different quantitative attributes : on one hand biological data (e.g. number of individuals for each taxon) and on the other hand physico-chemical data (e.g. pH, temperature, level of organic matter, level of dissolved oxygen, etc.). Taxons are characterized life traits that are qualitative data (e.g. lifetime of invertebrates).

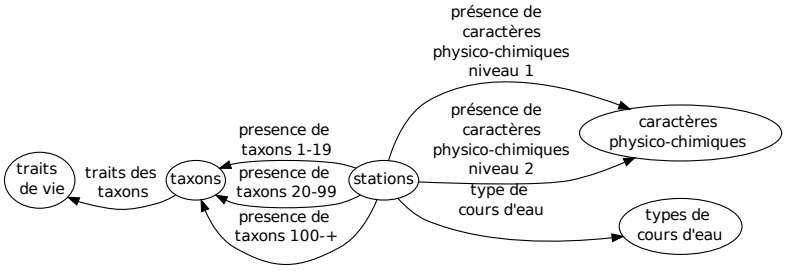

Fig. 5 Schema of data used by RCA.

Figure 4 presents data with a schema. Each node represents an object set which is translated in an object-attribute context. Each labeled edge represents a relation between object sets which is translated in an object-object context. For each edge we consider the relation in both direction.

We would like to extract from these data some relations between the different kind of information that describe a station. It would be interesting for instance to

extract rules with the format: *modality M of life trait T in a watercourse of type E implies the presence of physico-chemical character C*, which can be obtained by orienting thre relations of figure 4 such that the relation `taxons' traits` go from `taxons` to `life traits` and that all other relations have `stations` for domain as in figure 5.

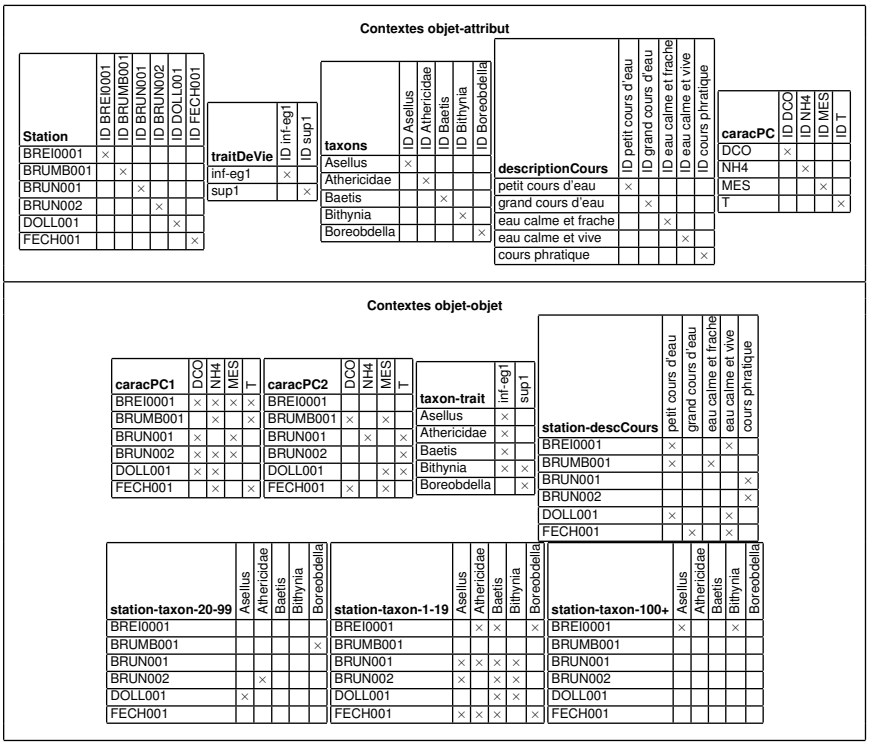

Table 4 Famille relationnelle de contextes obtenue partir de nos donnes d'exemples en considrant les directions de relation de la figure 5.

The Relational Context Family for this particular configuration is presented by table 4. We add in the object-attribute contexts an identifier as attribute for each object in order to create a concept for each object after the initialization step. Figure 6 presents needed excerpts of the lattice family obtained to extract the rule *Presence at medium level of a taxon which lifetime is more than one year implies a high Chemical oxygen demand.*[3] Indeed, we can notice that `Concept_3` is more specific that `Concept_41`. An attribute introduced by `Concept_3` implies an attribute introduced by `Concept_41` as all the attributes of `Concept_41` are inherited by `Concept_3`. `Concept_3` groups the stations that shelter a medium quantity of individuals from taxons grouped by `Concept_63`. `Concept_63` groups taxons

[3] considering the size of the example, this rule is to be considered for illustrative purpose only.

which individuals have a lifetime superior to one year. `Concept_41` groups the stations which physico-chemical characters (of level 2, i.e. a high level) is Chemical oxygen demand. From those observations we can induce the previous implication. From the complete lattice we can obtain the whole set of implication rules between life traits and physico-chemical characters by considering all the cases where physico-chemical characters are introduced by a concept and life traits are introduced by a more specific concept.

But rules in the following format can also be relevant: *modality M of life trait T can appear when the physico-chemical character C is present*. To generate these kind of rule, we need to change from the previous configuration the direction of the relations between `life traits` and `taxons` and between `taxons` and `stations`. There exists more configurations which results can be relevant and varying the scaling operators increases the expression of the rules obtained with RCA.

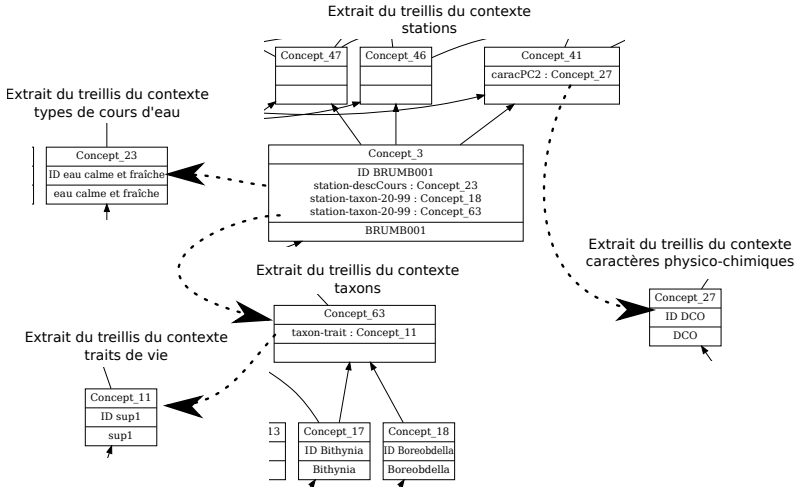

Fig. 6 Excerpts of lattices obtained from the relational context family of table 4.

If we consider the schema from figure 4 as a graph, the exploration consists in analyzing the different edges until we obtain a path between the physico-chemical characters and the life traits. By comparing the obtained results during an exploration with the results obtained using the classic RCA on the same relation, we notice that the final lattices are smaller and easier to read in the first case.

Combining the whole set of possible configurations is not to consider as by multiplying the relations we take the risk to raise a combinatorial explosion of the number of concepts, thus increasing the computation time and the complexity of the obtained concepts. For our example, considering all the relations in both directions scaled only by $\exists$ lead to the creation of 120 concepts against 66 and 63 concepts for the two configuration previously presented. So we plan an approach where the user

can explore different configuration by making different choices at each step of the process as presented in section 1.

5 Conclusion and discussion

In this article, we have presented an exploratory approach to assist the use of RCA in a way more appropriated for knowledge retrieval process. We have several motives to modify the original RCA process: to obtain relevant results faster by computing less lattices (preferably only the lattices that are of interest), to decrease the complexity of relational data mining, or to let the expert guide the discovery process based on its intuition and the learning patterns that appear along the process.

Several questions arise on this concept extraction approach from relational data. The initialization step strongly impact the discovered structures. It can speed up the process, if the object-object relations contain the information needed by the expert, or on the contrary it can hide relevant information to the expert. Nevertheless, the most important problem comes from the fact that modifications at each step implies that the concept generation is not monotonous anymore and it becomes possible to build examples where the process diverges by iterating on several recurring configurations.

In the original RCA process, when the fix point is reached the lattices on the two last step are isomorphic. So when a concept refers another one *via* a relational attribute the referenced concept can be found in a lattice from the same step. But with the exploratory process that we are proposing, when a concept refers to another, the referenced concept is in a lattice of the previous step and this concept can also refers to a concept from a lattice in another previous step. So we need to find solutions to present the information to the expert in a way simple enough to interpret. However, we think that such an exploratory approach is more applicable than a systematic approach that iterate until a fix point and give results too difficult for an expert to interpret.

References

[Azmeh et al(2011)Azmeh, Huchard, Napoli, Hacene, and Valtchev] Azmeh Z, Huchard M, Napoli A, Hacene MR, Valtchev P (2011) Querying relational concept lattices. In: Proc. of the 8th Intl. Conf. on Concept Lattices and their Applications (CLA'11), pp 377–392

[Barbut and Monjardet(1970)] Barbut M, Monjardet B (1970) Ordre et Classification: Algèbre et Combinatoire, vol 2. Hachette

[Bertaux et al(2009a)Bertaux, Le Ber, Braud, and Trémolières] Bertaux A, Le Ber F, Braud A, Trémolières M (2009a) Identifying ecological traits: a concrete fca-based approach. In: Ferré S, Rudolph S (eds) 7th International Conference on Formal Concept Analysis, ICFCA 2009, Darmstadt, Springer-Verlag, LNAI 5548, pp 224–236

[Bertaux et al(2009b)Bertaux, Le Ber, Li, and Trémolières] Bertaux A, Le Ber F, Li P, Trémolières M (2009b) Combiner treillis de Galois et analyse factorielle multiple pour

l'analyse de traits biologiques. In: d'Aubigny G (ed) Actes des XVIèmes Rencontres de la Société Francophone de Classification, Grenoble, pp 117–120

[Carpineto and Romano(2004)] Carpineto C, Romano G (2004) Concept Data Analysis: Theory and Applications. Wiley

[Dolques et al(2009)Dolques, Huchard, and Nebut] Dolques X, Huchard M, Nebut C (2009) From transformation traces to transformation rules: Assisting model driven engineering approach with formal concept analysis. In: Supplementary Proceedings of ICCS'09, pp 15–29

[Dolques et al(2010)Dolques, Huchard, Nebut, and Reitz] Dolques X, Huchard M, Nebut C, Reitz P (2010) Fixing generalization defects in UML use case diagrams. In: CLA'10: 7th International Conference on Concept Lattices and Their Applications, pp 247–258

[Ganter and Kuznetsov(2001)] Ganter B, Kuznetsov SO (2001) Pattern structures and their projections. In: Proc. of the 9th Int. Conf. on Conceptual Structures (ICCS 2001), pp 129–142

[Ganter and Wille(1999)] Ganter B, Wille R (1999) Formal Concept Analysis, Mathematical Foundations. Springer

[Huchard et al(2007)Huchard, Hacène, Roume, and Valtchev] Huchard M, Hacène MR, Roume C, Valtchev P (2007) Relational concept discovery in structured datasets. Ann Math Artif Intell 49(1-4):39–76

[Valtchev et al(2004)Valtchev, Missaoui, and Godin] Valtchev P, Missaoui R, Godin R (2004) Formal concept analysis for knowledge and data discovery : New challenges. In: Proc. of the 2nd Intl. Conf. on Formal Concept Analysis (ICFCA'04), pp 352–371