Soundness and Completeness of Relational Concept Analysis

Mohamed Rouane-Hacene[1], Marianne Huchard[2],
Amedeo Napoli[3], and Petko Valtchev[1]

[1] Dépt. d'Informatique UQAM, C.P. 8888, Succ. Centre-Ville, Montréal H3C 3P8, Canada
[2] LIRMM (CNRS - Université de Montpellier 2), 161 rue Ada, 34392 Montpellier - France
[3] LORIA UMR 7503, B.P. 239, 54506 Vandœuvre-lès-Nancy Cedex, France

Abstract. Relational Concept Analysis (RCA) is an extension of Formal Concept Analysis (FCA) to the processing of relational datasets, i.e., made of (objects $\times$ properties) contexts and (objects $\times$ objects) relations. RCA constructs a set of fixpoint concept lattices by iteratively expanding the lattices of the initial contexts. To that end, at each iteration a scaling mechanism translates the inter-object links into relational attributes that reflect the available conceptual structures. The output of a RCA task has so far only been described operationally. We propose here an analytic characterization thereof, i.e., a completeness and consistence result connecting fixpoint extents to particular relational structures in the input data.

1 Introduction

Formal Concept Analysis (FCA) [7] is a mathematical method that turns a set of individuals described by properties, called *formal context*, into a hierarchy of concepts (clusters of individuals and properties) that is a complete lattice. The concept lattice, the set of concepts provided with a specialization order, emphasizes commonalities in descriptions (by property sets). FCA has been successfully exploited as a framework for both data mining and knowledge discovery [6]. However, when realistic datasets are considered, the complex information available within the data, e.g., relational links, exceeds the computational power of classical FCA.

The processing of datasets described with a relational formalism (logic, graph-based, etc.) in FCA is an actively researched topic with many approaches reported in the literature [12, 15, 17]. Relational Concept Analysis (RCA) [8] has been proposed as an approach for mining potentially useful abstractions from relational data, e.g., roles that link concepts in ontology models. RCA input compares to multi-relational datasets whereas output is compatible with popular knowledge representation formalisms, the description logics (DL), a.k.a. DL-based languages [3].

The input data in RCA comprises a set of formal contexts, each corresponding to a sort of individuals, and a set of binary relations, each connecting the respective sets of individuals from two contexts. An RCA analysis task extracts a set of concept lattices, one per formal context, in a simultaneous and iterative way. Starting with the standard lattices of the initial contexts, the underlying construction method, called *Multi-FCA*, gradually translates the inter-object links into synthetic attributes, called *relational*, that

	Senior	Adult	Male	Female	Bleeding	Breathdisorder	Fatigue	Hairloss	Headache	HeartFailure	Hives	Nausea	Oedema	Vomiting
John		×	×		×	×	×			×		×		×
Carol		×		×	×		×	×			×		×	
Alex	×		×					×	×	×	×	×	×	
Mary		×		×			×	×					×	

Table 1. Context $\mathcal{K}_p$ encoding AIDS patients with their adverse reactions (ADR).

also reflect the available concepts at the links' destination objects. As adding new attributes to contexts typically extends the concept set, the processing is repeated until a fix-point of maximally extended contexts, with respective lattices, is reached. The output lattices highlighting all the cases of property sharing between objects, inclusive properties that are not owned by the objects themselves but by sets of other objects that are "connected" to the former ones through path-like structures of relational links.

The present paper sheds some light on the genesis of relational attributes. It investigates the way they encompass and extend one-valued FCA attributes and provide a necessary and sufficient condition for their formation. To that end, graph reasoning is applied to the network of objects and links induced by an RCA dataset.

The remainder of the paper is organized as follows: First, RCA is motivated against core FCA and a summary of the current knowledge about RCA is provided (Sec. 2). The analytical description of the RCA output is presented next (Sec. 3) followed by a survey of related work (Sec. 4) and concluding remarks (Sec. 5).

2 FCA on relational data

Key RCA structures and major results are summarized below (see details in [16]).

2.1 Basics of FCA

FCA is an order-theoretic approach suitable for knowledge discovery tasks as it abstracts concepts and conceptual hierarchies out of a collection of individuals described by properties. Core FCA encodes data in a *formal context* $\mathcal{K} = (O, A, I)$, where O is set of (formal) objects, A is set of (formal) attributes, and $I \subseteq O \times A$ is an incidence relation (comparable to a set of ground expressions $a(o)$, $a \in A$, $o \in O$). Its output is a complete lattice $\mathcal{L}$ made of all *(formal) concepts*, i.e., pairs of sets (X, Y) – a set of objects X (*extent*) and a set of attributes Y (*intent*) – such that all attributes in Y are shared by all objects in X and both sets are *maximal* w.r.t. this property. Henceforth, we shall use as a running example a medical dataset representing AIDS patients and the observed adverse reactions (ADR) to medication: Table 1 illustrates a formal context $\mathcal{K}_p$ while Figure 1 shows its concept lattice $\mathcal{L}_p$.

Beside the above one-valued attributes, FCA admits many-valued ones (e.g., the age of patients). Before processing such datasets, a.k.a. *many-valued contexts*, many-valued attributes are translated into one-valued by various *scaling* mechanisms.

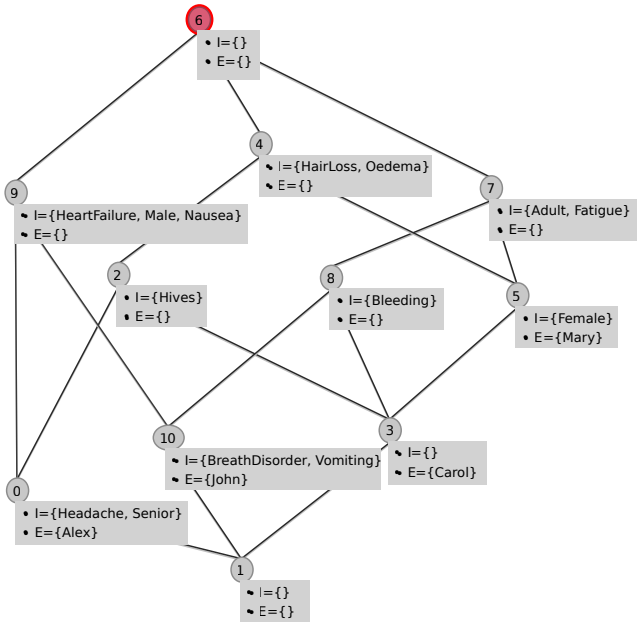

Fig. 1. The concept lattice $\mathcal{L}_p$ of the context shown in Table 1.

2.2 RCA, an approach for FCA of relational datasets

Relational datasets stem from eponymous databases and comply to the *Entity-Relationship* formalism. Thus, they are typically composed of several relational tables representing independent objects sorts (drugs, therapies, hospitals, etc.) or the relationships between those (e.g., patient-takes-drug). In FCA terms, each object sort can be hosted in a dedicated formal context. For instance, a HIV-centered pharmacovigilance[4] dataset would include, beside patient collection, a set of drugs. Moreover, like patients, these could be described by the underlying active agents as well as by their known ADR (see Tab. 2).

To express the *relational* information, i.e., the relationships among objects of the dataset, a collection of (objects $\times$ objects) binary *relations*[5] are added. In FCA, they can be conveniently represented as cross-tables: Tab. 3 shows the *patient-takes-drug* (left) and *drug-to-drug interaction* (right) relations.

[4] Pharmacovigilance is a bio-medical field monitoring the ADR to newly introduced drugs.
[5] Thus, higher-arity relations are excluded.

	Actinomycin	Efavirenz	Maraviroc	Raltegravir	Ritonavir	Tenofovir	Breathdisorder	Diarrhea	Fatigue	Hairloss	Headache	HeartFailure	LiverDamage	Nausea	Rash	Vomiting
Aluvia					×			×	×		×		×	×	×	×
Vicriviroc			×									×			×	
Truvada						×		×					×	×		×
Cosmegen	×						×		×	×						
Isentress				×				×			×			×		
Stocrin		×						×	×						×	×

Table 2. Context $\mathcal{K}_D$ of anti-HIV drugs with the expected ADR and active agents.

Hereafter, a set of mathematical notations will be used. First, the relations RCA admits are defined over pairs of object sets: i.e., each relation r is $r \subseteq A \times B$, where A and B are some predefined object sets (e.g., corresponding to the set O from a particular context). The latter are called the **domain** and the range of r, respectively (denoted $dom(r)$ and $ran(r)$). Next, for such a r, the set of r-successors of an object $o \in dom(r)$ w.r.t. r is $r(o) = \{\bar{o} \mid (o, \bar{o}) \in r\}$.

As an input data format for RCA, a unique structure, called *relational context family* (RCF), holds all the contexts and relations together.

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

- $\mathbf{K} = \{\mathcal{K}_i\}_{i=1,\ldots,n}$ *is a set of contexts* $\mathcal{K}_i = (O_i, A_i, I_i)$ *and*
- $\mathbf{R} = \{r_k\}_{k=1,\ldots,m}$ *is a set of relations* r_k *(*$r_k \subseteq O_j \times O_l$ *for some* $j, l \in \{1, \ldots, n\}$*).*

Associated with an RCF, a function rel maps a context $\mathcal{K} = (O, A, I) \in \mathbf{K}$ to the set of all relations r starting at its object set O: $rel(\mathcal{K}) = \{r \in \mathbf{R} \mid dom(r) = O\}$.

Our running example RCF is made of the contexts $\mathcal{K}_P$ (Tab. 1) and $\mathcal{K}_D$ (Tab. 2) and the relations `takes`, its inverse *is taken by* (`itb`) and *interacts with* (`iw`), shown in Tab. 3.

2.3 Turning relational links into first-class attributes

In dealing with relations from a RCF, i.e., the directed links between objects, RCA follows an approach which amounts to "propositionalizing" [9] them. In short, the links are translated into one-valued attributes that are further assigned to the objects at their origins. Since the mechanism compares to FCA scaling, we called it *relational scaling*[6].

The syntax and the semantics of the resulting *relational attributes* have been inspired by *role restrictions* of the DL formalism [3]: given a relation $r \subseteq O_i \times O_j$ and object $o \in O_i$, to assign a relational attribute to o, the set of its r-successors $r(o)$ is matched against a set of objects from O_j. The latter is typically the extent of a concept

[6] The term was first used in [15], with substantially different meaning.

takes	Cosmegen	Isentress	Aluvia	Vicriviroc	Stocrin	Truvada
Alex			×			×
Carol				×	×	
Mary			×			×
John	×	×				

interacts with	Cosmegen	Isentress	Aluvia	Vicriviroc	Stocrin	Truvada
Cosmegen						
Isentress						
Aluvia				×	×	
Vicriviroc			×			
Stocrin			×			×
Truvada					×	

Table 3. Left: Binary relation takes linking AIDS patients to anti-HIV drugs. The relation is taken by (henceforth referred to as itb) is the inverse of takes, *i.e.* ($\texttt{takes}^{-1}$). **Right**: The binary relation interacts with (iw) models interactions among drugs.

c over O_j, but could be any *named* set of objects. The overall pattern for naming the attributes is q r : c where q is a *quantifier*, r is the relation and c the identifier (here a concept name) of an object set $X \subseteq O_j$.

The exact matching discipline for $r(o)$ and X depends on q which, for the current study, is chosen within the set $\mathbf{Q} = \{\forall, \exists, \forall\exists, \geq, \geq_q, \leq, \leq_q\}$. The possible disciplines are schematized by a generic function, κ, whose effect is to filter the objects from O_i to get an attribute $q\ r : c$. Formally, given a relation r and a quantifier q, κ maps an object set from $ran(r)$ to an object set from $dom(r)$:

$$\kappa : \mathbf{Q} \times \mathbf{R} \times \bigcup_{i=1,\ldots,n} \wp(O_i) \rightarrow \bigcup_{i=1,\ldots,n} \wp(O_i).$$

Its instanciations w.r.t. to the quantifiers q are provided in Tab. 4 (columns three and four). For instance, consider the concept lattice in Fig. 1 and its concept c_9. The extent is $X = \{\texttt{Alex}, \texttt{John}\}$. With the existential quantification operator, $\kappa(\exists, \texttt{itb}, X)$ is the set of drugs taken by at least one patient from X ($\{\texttt{Aluvia}, \texttt{Cosmegen}, \texttt{Isentress}, \texttt{Truvada}\}$).

Operator name	Notation	Attribute template	$\kappa(q, r, Ext(c))$ calculation
Universal (wide)	$\mathbb{S}_{(r,\forall),\mathcal{B}}$	$\forall\ r : c$	$r(o) \subseteq Ext(c)$
Existential	$\mathbb{S}_{(r,\exists),\mathcal{B}}$	$\exists\ r : c$	$r(o) \cap Ext(c) \neq \emptyset$
Universal strict	$\mathbb{S}_{(r,\forall\exists),\mathcal{B}}$	$\forall\exists\ r : c$	$r(o) \subseteq Ext(c), r(o) \neq \emptyset$
Cardinality restriction (max)	$\mathbb{S}_{(r,\geq),\mathcal{B}}$	$\geq\ n\ r : \top_{\mathcal{L}}$	$\lvert r(o)\rvert \geq\ n$
Cardinality restriction (min)	$\mathbb{S}_{(r,\leq),\mathcal{B}}$	$\leq\ n\ r : \top_{\mathcal{L}}$	$\lvert r(o)\rvert \leq\ n$
Qualified card. restriction (max)	$\mathbb{S}_{(r,\geq_q),\mathcal{B}}$	$\geq\ n\ r : c$	$r(o) \subseteq Ext(c), \lvert r(o)\rvert \geq\ n$
Qualified card. restriction (min)	$\mathbb{S}_{(r,\leq_q),\mathcal{B}}$	$\leq\ n\ r : c$	$r(o) \subseteq Ext(c), \lvert r(o)\rvert \leq\ n$

Table 4. Relational scaling operators in RCA: names, notations, and produced attributes with incident object sets ($Ext(c)$ is the extent of a concept c).

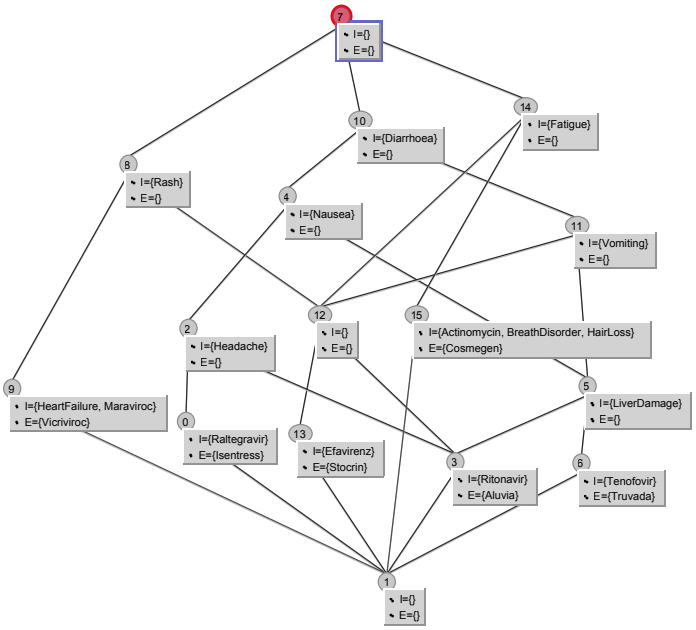

Fig. 2. The concept lattice $\mathcal{L}_D$ of the context shown in Table 2.

Obviously, the κ can be applied to a family of sets $\mathcal{B}$ over $ran(r)$, in particular, the entire set of concept extents from a given concept lattice $\mathcal{L}$. This is the motivation behind the definition of context-level scaling operators $\mathbb{S}_{(r,q),\mathcal{B}}$ (column two from Tab. 4). The following definition provides a general pattern for such operators specifying the way the generated attributes expand the basic attribute set of the argument context:

Definition 2 (Scaling operator $\mathbb{S}_{(r,q),\mathcal{L}}$)
Given a context $\mathcal{K}_i = (O_i, A_i, I_i)$ and a relation $r \in rel(\mathcal{K}_i)$, with $ran(r) = O_j$, let $\mathcal{L}_j$ be a concept lattice over O_j. The scaling operator $\mathbb{S}_{(r,q),\mathcal{L}_j}$ over $\mathcal{K}_i$ yields the derived context $(O^+, A^+, I^+) = \mathbb{S}_{(r,q),\mathcal{L}_j}(\mathcal{K}_i)$, where:

- $O^+ = O$,
- $A^+ = \{`q\ r : c' \mid c \in \mathcal{L}_j\}$,
- $I^+ = \bigcup_{c \in \mathcal{L}_j} (\kappa(q, r, Ext(c)) \times \{`q\ r : c'\})$.

Tab. 5 illustrates $\mathbb{S}_{(\texttt{itb},\exists),\mathcal{L}_P}(\mathcal{K}_D)$, the result of scaling the drug context $\mathcal{K}_D$ (Tab. 2) along `itb` with an existential operator upon the lattice $\mathcal{L}_P$ (Fig. 1).

The next step in transforming the relational information about the objects from $\mathcal{K}_i$ is to scale upon every relation in $rel(\mathcal{K}_i)$ and then to append the results to $\mathcal{K}_i$. To that end, we define a function $\rho : \mathbf{R} \to \mathbf{Q}$ that maps[7] relations to scaling operators from $\mathbf{Q}$.

[7] A non functional ρ, albeit plausible, was willingly excluded for simplicity reasons.

	∃ *itb:c0*	∃ *itb:c2*	∃ *itb:c3*	∃ *itb:c4*	∃ *itb:c5*	∃ *itb:c6*	∃ *itb:c7*	∃ *itb:c8*	∃ *itb:c9*	∃ *itb:c10*
Aluvia	×	×		×	×	×	×		×	
Vicriviroc		×	×	×	×	×	×	×		
Truvada	×	×		×	×	×	×		×	
Cosmegen						×	×	×	×	×
Isentress						×	×	×	×	×
Stocrin		×	×	×	×	×	×	×		

Table 5. The existential scaling of the drug context $\mathcal{K}_D$ along the relation itb using the lattice of AIDS patients. Observe that $\exists$ *itb:*c_1 is skipped as c_1 is the bottom concept whose extent is void.

Let $\mathbf{L}$ be the set of lattices corresponding to contexts from $\mathbf{K}$. Assume now $\mathcal{K} \in \mathbf{K}$ with $rel(\mathcal{K}) = \{r_l\}_{l=1,\ldots,m_\mathcal{K}}$ and, let for each r_l $O_{j_l} = ran(r_l)$ with $\mathcal{L}_{j_l} \in \mathbf{L}$ being the lattice on O_{i_l}. Now, the complete relational extension of $\mathcal{K}$ with respect to ρ and $\mathbf{L}$, denoted $\mathbb{E}_{\rho,\mathbf{L}}$, is the *apposition* [7] of $\mathcal{K}$ with the respective derived context yielded by scaling upon each r_l with its $\rho(r_l)$:

Definition 3 (Complete relational extension of a context)
Given a RCF $(\mathbf{K}, \mathbf{R})$*, with a set of lattices* $\mathbf{L}$*, a scaling operator mapping* ρ*, and a context* $\mathcal{K} \in \mathbf{K}$ *with* $rel(\mathcal{K}) = \{r_l\}_{l=1,\ldots,m_\mathcal{K}}$*, the complete relational extension of* $\mathcal{K}$ *w.r.t.* ρ *and* $\mathbf{L}$ *is*

$$\mathbb{E}_{\rho,\mathbf{L}}(\mathcal{K}) = \mathcal{K} \mid \mathbb{S}_{(r_1,\rho(r_1)),\mathcal{L}_{i_1}}(\mathcal{K}) \mid \ \ldots \mid \mathbb{S}_{(r_{m_k},\rho(r_{m_k})),\mathcal{L}_{i_{m_k}}}(\mathcal{K})$$

Let now $\mathcal{K}^e = \mathbb{E}_{\rho,\mathbf{L}}(\mathcal{K})$ is the complete relational extension for some $\mathcal{K} \in \mathbf{K}$. Both the original and the extended context share the same object set, yet $\mathcal{K}^e$ has a larger set of attributes hence a larger lattice. Indeed, following Lemma 2 from [16], its lattice $\mathcal{L}^e$ comprises all the extents from the lattice $\mathcal{L}$ of $\mathcal{K}$, plus possibly some additional ones (a general rule with apposed contexts [19]).

Now the application of the complete relational extension operator $\mathbb{E}_{\rho,\mathbf{L}}$ to all contexts from $\mathbf{K}$ yields a set operator $\mathbb{E}^*_{\rho,\mathbf{L}}$ over $\mathbf{K}$: The resulting $\mathbf{K}^e = \mathbb{E}^*_{\rho,\mathbf{L}}(\mathbf{K})$ is made of all the $\mathcal{K}^e$ and, correspondingly, its lattice set $\mathbf{L}^e$ comprises all $\mathcal{L}^e$. The immediate consequence thereof is that $\mathbf{L}^e$, while preserving the concepts from $\mathbf{L}$, may include some additional ones, hence it represents a finer conceptualization of the RCF data. This in turn warrants a new scaling step $\mathbb{E}^*_{\rho,\mathbf{L}^e}(\mathbf{K}^e)$ that may, in turn, effectively extend the set of available attributes and hence, once more, generate previously unseen concepts.

In summary, the overall process of analyzing an RCF can be schematized as an iterative application of $\mathbb{E}^*_{\rho,\mathbf{L}^e}$ to the initial set of contexts from the RCF. The underlying analysis method is presented below.

2.4 Iterative lattice construction

RCA constructs a concept lattice for each $\mathcal{K}_i$ starting with the lattice $\mathcal{L}_i$ built with the original attribute set A_i. At subsequent steps, it alternates (i) generation of relational

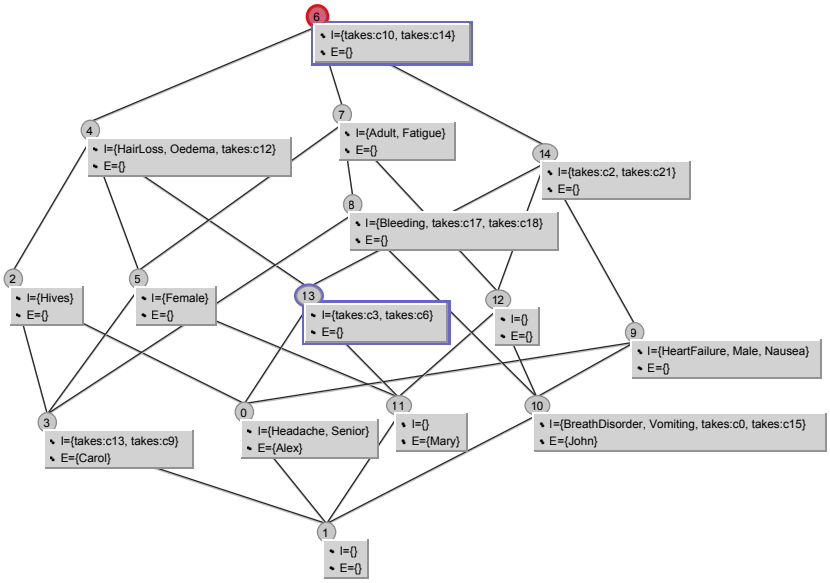

Fig. 3. The final lattice of patients ($\mathcal{L}_P^{\infty}$). Quantifiers are omitted in relational attributes because of visualization limitations of GALICIA.

attributes by relational scaling with concepts discovered at *the previous iteration*, and (ii) lattice maintenance, i.e., the expansion of the current concept lattice with the newly synthesized attributes. As shown in [16], the process *converges*, i.e., from a particular iteration onward, no new concepts emerge in $\mathbf{L}^e$, hence the scaling step yields no new attributes and the whole process halts. Algorithm 1 puts that into pseudo-code.

Spelled differently, the computation stabilizes at a global fixpoint represented by the set of contexts and their lattices. Yet no analytical description has been provided so far for the fixpoint lattice family w.r.t. the initial data in the RCF.

To study the fixpoint structures, we capture the way MULTI-FCA operates in the definition of a sequence of *non contracting* contexts. By *non contracting* it is meant contexts whose respective components either grow or remain stable. Indeed, in our case, each relationally-extended version of a context has the same object set, yet potentially bigger attribute set and hence incidence relation. The respective lattices follow the same trend: each extended version has the same extent family plus potentially some new object sets as extents. Yet the size of the lattice is bounded by $2^{|O|}$, hence new concepts cannot be created *ad infinitum*.

Formally, each context $\mathcal{K}_i \in \mathbf{K}$ from the input RCF yields a sequence $\mathcal{K}_i^p$ whose zero member $\mathcal{K}_i^0 = (O_i^0, A_i^0, I_i^0)$ is the input context $\mathcal{K}_i$ itself. From there on, each subsequent member is the complete relational expansion of the previous one w.r.t. ρ and the lattices of the previous iteration. This yields a global sequence of context sets $\mathbf{K}^p$ and the corresponding sequence of lattice sets $\mathbf{L}^p$.

```
1: proc MULTI-FCA(
2: In: (K, R) an RCF, ρ an operator mapping
3: Out: L a set of lattices)
4: p ← 0 ; halt ← false
5: for i from 1 to n do
6:    K_i^0 ← SCALE(K_i)
7:    L_i^0 ← BUILD-LATTICE(K_i^0)
8: while not halt do
9:    p = p + 1
10:   for i from 1 to n do
11:      K_i^p ← EXTEND-CONTEXT(K_i^{p-1}, ρ, L^{p-1})
12:      L_i^p ← UPDATE-LATTICE(K_i^p, L_i^{p-1})
13:   halt ← ∀i ∈ {1,...,n}, ISOMORPHIC(L_i^p, L_i^{p-1})
```

Algorithm 1: Producing a lattice for each context in an RCF.

Definition 4 *Given a* RCF $(\mathbf{K}, \mathbf{R})$ *and a scaling operator mapping* ρ*, the sequence of context sets* $(\mathbf{K}^j)_{j\in\mathbb{N}}$ *is recursively defined as*

$$\mathbf{K}^0 = \mathbf{K} \ ; \ \mathbf{K}^{p+1} = \mathbb{E}^*_{\rho,\mathbf{L}^p}(\mathbf{K}^p)$$

In [16] it is shown that each $\mathcal{K}_i^p$ as well as the entire $\mathbf{K}^p$ are non-contracting while naturally bounded from above (by the bounded sizes of the lattices in $\mathbf{L}^p$). Hence, all sequences converge toward their respective limits.

Theorem 1 *Given a* RCF $(\mathbf{K}, \mathbf{R})$ *and a scaling operator mapping* ρ*, the sequence* $(\mathbf{K}^p)$ *converges towards a well-defined set of maximally extended contexts* $\mathbf{K}^\infty$.

As shown in Algorithm 1, the test for $\mathbf{K}^\infty$ succeeds whenever a p is reached s.t. $\mathbb{E}^*_{\rho,\mathbf{L}^p}$ produces no new concepts at any of the contexts. The fixpoint lattices of our pharmacovigilance dataset are given in Fig. 4 and Fig. 3. Obviously, the fixpoint depends on ρ: it is conceivable that the same RCF yields a different outcome for another combination of quantifiers.

RCA has been implemented in GALICIA [18, 1] and is currently operational for various applications, such as reengineering of software models [4], refactoring of object-oriented code [13], etc.

3 Soundness and completeness of the MULTI-FCA method

3.1 Observations on the iterative analysis process

The iterative analysis process ends up with a collection of lattices whose concept intents mix attributes from the input RCF and relational ones created by scaling. While the former admit straightforward interpretation, the latter have more complex semantics and may prove hard to interpret, especially in large concept intents. We therefore need to clarify the semantics of the expressions found in fixpoint intents, e.g., $\{\texttt{Rash}, \texttt{itb:c2}, \texttt{itb:c5}, \texttt{iw:c12}, \texttt{itb:c7}\}$ in concept c_8 of $\mathcal{L}_D^\infty$ (Fig. 4). The question to ask is: What exactly do these expressions say about the *initial* RCF data?

Intuitively, the relational attributes in a fixpoint contexts $\mathcal{K}_i^\infty$ are all rooted in the initial set A_i^0. However, the exact connection is blurred by a number of iterations, using scaling and arbitrary combinations to form intents. Therefore, to successfully ground the interpretation of the RCA output, we need formally established results on:

- the *nature* of configurations in the data (e.g., graphs, trees, sequences of inter-linked objects) that are reflected in each fixpoint concept,
- the *correctness* of the iterative method: only concepts mirroring that sort of structures are generated (absence of spurious concepts in the output),
- the *completeness* of the method: no relevant structure in the data is left unrepresented in the final result (exhaustiveness of the set of generated concepts).

In short, we face a language whose expressions must be provided with clear semantics. They can only be denotational semantics: As we observed above, O_i remain unchanged all along the analysis process whereas the discovered concept extents never vanish in the iterative process which means the concept refinement is *monotonic*.

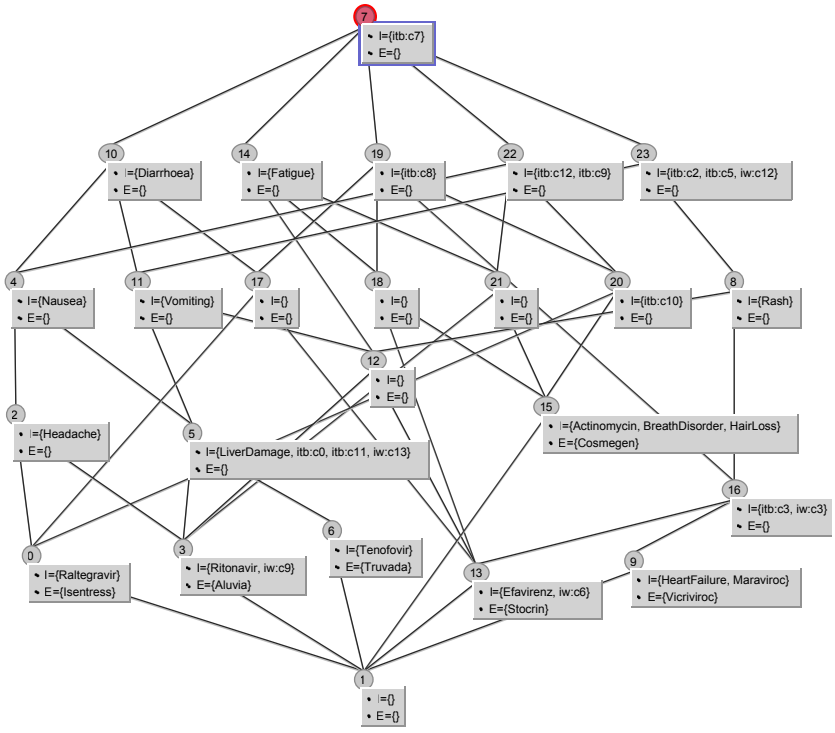

Fig. 4. The final lattice of drugs ($\mathcal{L}_D^\infty$). Quantifiers are omitted in relational attributes due to visualization limitation of GALICIA.

As a first step toward a more comprehensive answer to the semantic question, we define below a graph-like structure on families of object sets. Its immediate goal is to

"explain" the genesis of attributes and intents in the fixpoint contexts by tracing their links back to A_i^0. The structures reflect two types of relationships:

- *scaling*: a concept extent yields an attribute extent,
- *generation*: set of attribute extents combine into a concept extent (through $\cap$).

Clearly, the target structures in the input data depend on the ρ function.

3.2 Basic definitions and notations

To focus on the extents of a context $\mathcal{K}_i$ while ignoring the remaining object sets that are irrelevant, we introduce the notion of *image*. Images differ by the nature of the generating attribute set (single attributes vs. multiple ones) and order (depth in the structure induced by the links of the above two types).

Definition 5 (Images, atomic and compound)
Given a context $\mathcal{K}_i = (O_i, A_i, I_i)$, a set $X \subseteq O_i$ is:

atomic image (AI) *if $\exists a \in A_i$ s.t. $X = a'$,*
compound image (CI) *if $\exists J_X \subseteq \mathbb{N}$ and AIs $\{X_j\}_{j \in J_X}$ s.t. $X = \bigcap_{j \in J_X} X_j$.*

For instance, the set $\{\texttt{John}\}$ is an AI (and thus a CI) whereas $\{\texttt{Carol}\}$ is a CI but not an AI. Clearly, CIs correspond to concept extents in the initial contexts $\mathcal{K}_i^0$.

To distinguish the images generated by the initial attribute sets A_i^0 that are the basis of the entire generation process from those in the scaled contexts, we split the overall set into orders. Thus, the images in $\mathcal{K}_i^0$ are qualified as 0-order AI/CI, shortened to 0-AI/0-CI. Images from scaled contexts at different steps of the iterative process typically have higher orders. These are defined recursively:

Definition 6 (k-order images, atomic and compound)
Given RCF $(\mathbf{K}, \mathbf{R})$*, ρ and $\mathcal{K}_i = (O_i, A_i, I_i)$ from $\mathbf{K}$, a set $X \subseteq O_i$ is:*

***k+1*-order atomic image (*k+1*-AI)** *if*
1. *X is not a p-CI for any $p \leq k$, and*
2. *$\exists r \in rel(\mathcal{K}_i)$ and $\exists Z \subseteq ran(r)$ which is a k-CI s.t. $X = \kappa_e(\rho(r), r, Z)$,*

***k+1*-order compound image (*k+1*-CI)** *if*
1. *$\exists J_X \subseteq \mathbb{N}$ and $\{Z_j\}_{j \in J_X}$ where each Z_j is a p_j-AI for some $p_j \leq$ k+1 s.t. $X = \bigcap_{j \in J_X} Z_j$, and*
2. *k+1 is minimal for that property, i.e., no such index set J_X for smaller values (k, k-1, etc.).*

For instance, the patient set $\{\texttt{Alex}, \texttt{John}, \texttt{Mary}\}$ is a 1-AI, whereas $\{\texttt{John}, \texttt{Mary}\}$ is a 1-CI that is not a 1-AI. Indeed, $\{\texttt{Alex}, \texttt{John}, \texttt{Mary}\} = \kappa_e(\exists, \texttt{takes}, \{\texttt{Aluvia}, \texttt{Isentress}\})$ whereas the the latter set of drugs is a *0*-AI since $\{\texttt{Aluvia}, \texttt{Isentress}\} = \{\texttt{Headaches}\}'$. Moreover, $\{\texttt{Alex}, \texttt{John}, \texttt{Mary}\}$ together with $\{\texttt{Carol}, \texttt{John}, \texttt{Mary}\} = \{\texttt{Adult}\}'$ contribute to the canonical generation of $\{\texttt{John}, \texttt{Mary}\}$. This situation is illustrated in Fig. 5 where both types of links are clearly distinguished.

The above definition basically says that in the global graph where CIs/AIs of various orders are connected by *generation* and *scaling* links, k-CIs require a minimal chain of

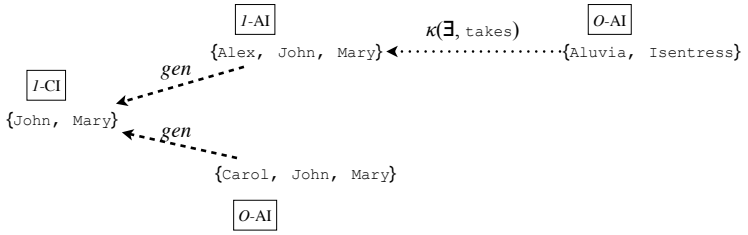

Fig. 5. An illustration of the genesis of the *1*-CI $\{\texttt{Mary}, \texttt{John}\}$.

$k+1$ generation/scaling links in order to emerge from the level-*0* CIs/AIs. The graph structure is easily shown to be a DAG. Moreover, observe that for X to be *k*-CI, at least one Z_j must be a *k*-AI.

Property 1 *If X is k-CI, then $\forall J_X \subseteq \mathbb{N}$, $\{Z_j\}_{j\in J_X}$ s.t. $X = \bigcap_{j\in J_X} Z_j$, $\exists j^* \in J_X$ with Z_{j^*} being a p-AI where $p \geq k$.*

In the following we shall provide a one-to-one mapping of fixpoint concept extents to *k*-CI for $k \in \{0, \ldots, t\}$ where t is the number of steps before termination in the iterative analysis process.

3.3 Correctness

Below, we show that with t steps before termination, every extent of a concept that is first created at step p, $p \leq t$ is in fact a *p*-CI. We start by providing some auxiliary definitions.

First, as we reason about the process output, w.n.l.g. we can assume that each attribute is assigned a unique *rank*. The rank is an integer number corresponding to the order of creation (by scaling) within the total set of fixpoint attributes: $rank : \bigcap_{i=1}^{n} A_i^{\infty} \rightarrow \mathbb{N}$. For the ranks of the initial attributes –that predate any scaling– we assume they are assigned in a way consistent with the above condition: their ranks represent a commencing segment of $\mathbb{N}$. For instance, in our RCF, we may assume that initial attribute ranks follow the left-to-right column order from the context tables (Tab. 1 and 2) with patient attributes coming before drug ones. Furthermore, the relational attributes follow the natural order of their names[8]. Thus, `itb:c0` has the lowest-rank among (31) them and `takes:c23` the highest (84).

Furthermore, based on attribute ranks, we assume a total order on arbitrary attribute sets which is the opposite of the standard string order –highest ranks are compared first– hence it is called *anti-alphabetic* (denoted $\leq_{a2}$). Formally, assume $Y_1, Y_2 \subseteq A_i^{\infty}$:

$$Y_1 \leq_{a2} Y_2 \textbf{ iff } \operatorname{argmax}(\{rank(a) \mid a \in Y_1 \triangle Y_2\}) \in Y_2.$$

Thus, $\{\texttt{Adult}, \texttt{Fatigue}, \texttt{takes:c10}, \texttt{takes:c14}\} \leq_{a2} \{\texttt{takes:2}, \texttt{takes:21}, \texttt{takes:c10}, \texttt{takes:c14}\}$ (the intents of patient concepts c_7 and c_{14}, respectively). It is readily shown that $\leq_{a2}$

[8] This only makes sense since the fixpoint is reached after a single scaling step.

is compatible with set-theoretic inclusion: For any $Y_1, Y_2 \subseteq A_i^\infty$, $Y_1 \subseteq Y_2$ entails $Y_1 \leq_{a2} Y_2$.

We also extend the notion of *generator* for a set of attributes to object sets $X \in O_i$: $Y \subseteq A_i^\infty$ is a generator of X whenever $X = \bigcap_{a \in Y} a'$. Now the *canonical generator* of X, $can(X)$ is the unique minimal one w.r.t. $\leq_{a2}$. It is readily shown that $can(X)$ is also minimal for set-theoretic inclusion. For instance, $can(\{\texttt{John}, \texttt{Mary}\}) = \{\texttt{Adult}, \texttt{takes:2}\}$.

Finally, attribute ranks are also expanded to sets of attributes and sets of objects. For a set $Y \subseteq A_i^\infty$, the ranks is the maximal of all member ranks: $rank(Y) = max(\{rank(a) \mid a \in Y\})$. In contrast, for $X \in O_i$, its rank is the canonical generator rank: $rank^o(X) = rank(can(X))$. Thus, $rank^o(\{\texttt{John}, \texttt{Mary}\}) = 63$. We can now formulate our first key result:

Theorem 2 *Given an* RCF $(\mathbf{K}, \mathbf{R})$*, a function ρ and a context $\mathcal{K}_i^\infty$ from $\mathbf{K}^\infty$, let $X \subseteq O_i$. In order for X to be generated as concept extent at step $p \leq t$ of the analysis process, it is necessary that X be a p-CI.*

Sketch of a proof Induction upon $rank^o()$: First, all extents X whose ranks $rank^o(X)$ are less or equal the highest rank of an initial attribute say s_a, clearly possess a generating set made exclusively of initial attributes. Hence X can be represented as an intersection of *0*-AI and therefore X is a *0*-CI. For X of ranks above s_a, say $rank^o(X) =$ *v+1*, we assume that for all extents Z of ranks $v_Z \leq v$, the conditions of the theorem are met (being created at step p_Z, a Z is a p_Z-CI). Using the attributes a from the canonical generator $can(X)$, X is further decomposed as an intersection of $Z_j = a'$ for $j \in J_X$, all of whom are created at steps $p_j \leq p$. By the inductive hypothesis, each Z_j is a p_j-CI and this provides the demonstration of X being a p-CI. As a special case, consider $can(X) = \{a\}$ with $a := \text{`}\rho(r)r : c'$ for some $r \in rel(\mathcal{K}_i)$ and some concept $c = (T, Y)$ over the objects in $ran(r)$. In this case $X = \kappa(\rho(r), r, T)$. Since T should already exist at step p, for X to be generated, its rank is at most v. Moreover, T can only be created at step *p-1*, hence by the inductive hypothesis it is a *p-1*-CI, which makes X a p-AI and hence p-CI. □

3.4 Completeness

We now tackle the opposite direction of the mapping, i.e., the proof of each p-CI being a concept extent in a fixpoint context.

Theorem 3 *Given an* RCF $(\mathbf{K}, \mathbf{R})$*, a function ρ and a context $\mathcal{K}_i^\infty$ from $\mathbf{K}^\infty$, let $X \subseteq O_i$. In order for X to be generated as concept extent by the analysis process, it is sufficient that there be a $p \in \mathbb{N}$ s.t. X is a p-CI.*

Sketch of a proof We use complete induction on p. In the base case $p = 0$, the proof is immediate following the Definition 5 (0-AI and 0-CI). Whether 0-AI or not, X is a 0-CI and as such is a concept intent.

Now the inductive hypothesis is for all $p \leq k$, X is k-CI entails X is a concept extent formed at the k-th step of the global iterative process In the inductive step, let $p =$ *k+1* and observe that by Definition 6, X is not p-CI for any $p \leq k$ (*). The reasoning now splits into complementary cases:

case 1 X is a *k+1*-AI. Thus, $\exists r \in rel(\mathcal{K})$ and $\exists \bar{X} \subseteq ran(r)$, $\bar{X}$ being a k-CI s.t. $X = \kappa(\rho(r), r, \bar{X})$. By the inductive hypothesis, $\bar{X}$ is an extent of a concept $c = (\bar{X}, \bar{Y})$ over the set $ran(r)$ generated at the k-th step of the process. From Definition 2, there will be an attribute $a_X := \text{`}\rho(r)r : c'$ in the scaled set A^{k+1} such that X is the image of a_X in O ($a'_X = X$). Consequently, there will be a concept $c_X = (X, Y)$ over O at that step. Now assuming c_X (hence X) was generated at an earlier step $p \leq k$ we show a contradiction: by Theorem 2, X is a p-CI, yet this is a contradiction with (*).

case 2 X is a *k+1*-CI. Thus, $\exists J_X \subseteq \mathbb{N}$ and $\{Z_j\}_{j \in J_X}$ s.t. $X = \bigcap_{j \in J_X} Z_j$ and each Z_j is a p-AI with $p \leq$ *k+1* and no such set J_X for smaller values exists. From Property 1 we know $\exists j^* \in J_X$ s.t. Z_{j^*} is a *k+1*-AI. Furthermore, from the above *case 1* of this proof, it follows that all such Z_{j^*} are the extents of relational attributes a created at step *k+1*. Now, from the inductive hypothesis, we know that the remaining Z_j are attribute extents created at earlier steps. However, this only says that X is generated –at latest– at step *k+1*. Thus, to formally demonstrate that X could not be generated at a step $p \leq k$, we assume the opposite and prove the contradiction: Assume now X, albeit a *k+1*-CI, is an extent generated at step $s \leq k$. Following Theorem 2, we deduce that X is also a s-CI. Hence $\exists J_s \subseteq \mathbb{N}$ and $\{Z_j\}_{j \in J_s}$ s.t. $X = \bigcap_{j \in J_s} Z_j$ and each Z_j is a p-AI with $p \leq s$. Yet this formally contradicts the fact that the above set J_X does not exist for values strictly less than $k + 1$. □

In summary, all concept extents in the fixpoint lattices are related to the original attribute extents by chains of links having one of the above two types. The critical chains that "explain" the genesis of a fixpoint extent X clearly pass trough its canonical generator or, more precisely, the extents of the member attributes. Furthermore, all such extents can be connected to their own canonical generators and so forth, all the way down to *0*-AIs. While we focused here on establishing the links between any such X and its multiple generating sets, additional work will be necessary to determine whether a single path in the overall graph can be associated to X. Beside providing a *canonical path* for an extent, this would also enable a more satisfactory answer to the natural question termination question, i.e., how many steps would MULI-FCA need to reach its fixpoint for a particular RCF.

4 Related work

RCA relates to approaches for extending the output of FCA towards relational expressions such as logical formulae, graphs, etc. For instance, in [14], the many-valued attributes are scaled upon a *fixed* hierarchy of concepts (the terms in the *TBox* of a DL knowledge base). In RCA terms, the method employs a single-step static relational scaling. Moreover, the approach critically depends on the availability of a suitable *TBox*. Simultaneously, the relations in FCA have been formalized as *power context families* (PCF) [15] where inter-object links (object pairs) are first-class formal objects of dedicated contexts. Yet the use of the corresponding concepts on links as descriptors for concepts on true objects, i.e., entities, remains unclear.

Independently, FCA has been explored as a tool for structuring DL knowledge bases. In [2], an FCA-based method constructs the hierarchy of all conjunctions of concepts

from a *TBox*. As it is centered on human-guided *attribute exploration*, the possible references between concepts are ignored. The *relational exploration* [17] expands the former method towards a full set of DL constructors, i.e., a target language closer to the one in RCA. As the method explores the syntactic structure of the DL formulae it fails to capture the existing references between the underlying DL concepts (e.g., via a subformula). For similar reasons, the generation of DL expressions needs to be restricted by some syntactic criteria (e.g., the depth of constructor nesting) since otherwise unbounded. A comparable generation mechanism for a richer DL language (e.g., inclusive disjunction), albeit without the closedness requirement on the produced descriptions, was explored in the *machine learning* system DL-LEARNER [11]. While producing concept descriptions structurally richer than the RCA output (strictly conjunctive), the system presents the same shortages as above, in particular, no recognition of the references between the discovered concepts. Back to FCA, in [5] a larger set of relational structures have been explored for concept construction yet with the same syntax-based techniques. Again, the generation of concept descriptions is controlled by limiting the nesting depth.

In a parallel trend, graph-based descriptions of the formal objects are assumed [10, 12]). Despite the broad coverage of the graph-based formats, e.g., chemical compound models or social networks, the proposed methods are not suitable for generating DL-like concept descriptions. Indeed, such methods are inherently limited to graph-like concept descriptions (e.g., no quantifiers) with only intra-object relations, i.e., relations among the parts of whole (as in chemical models).

5 Conclusion

RCA is an analysis framework that extends core FCA to the processing of relational datasets, i.e., with multiple sorts of objects and relational links between them. It thus constructs a set of lattices, one per object sort. The associated method performs a special kind of propositionalization on the links, a relational scaling, that yields standard FCA attributes with various semantics borrowed from the DL formalisms. It iterates upon the initial data, swapping at each iteration, the scaling with the maintenance of the current lattices until a fixpoint is reached.

Here we presented the framework and analyzed its *modus operandi* in order to provide an analytic characterization of the fixpoint lattice set. To that end, we defined images, the equivalent of concept extents, without the ambiguity of the multiple generations. A major advantage thereof is that they can be easily traced back to the initial data. We demonstrated the equivalence between fixpoint extents and higher-order images in two separate theorems that establish the correctness and completeness of our method, respectively.

Having established the theoretical foundations of RCA, our next major concern will be to make it a practical tool. To that end we shall focus on performances and study alternative techniques for speeding up the computing of updated lattices at subsequent iterations. In this respect, challenging, and still open, question is how to properly estimate the number of iterations RCA would require on a particular dataset.

References

1. http://sourceforge.net/projects/galicia/.
2. F. Baader. Computing a minimal representation of the subsumption lattice of all conjunctions of concepts defined in a terminology. In *Proc. of the Intl. Symp. on Knowledge Retrieval, Use, and Storage for Efficiency (KRUSE'95)*, pages 168–178, Santa Cruz, USA, 1995.
3. F. Baader, D. Calvanese, D. McGuinness, D. Nardi, and P. Patel-Schneider, editors. *The Description Logic Handbook*. Cambridge University Press, Cambridge, UK, 2003.
4. M. Dao, M. Huchard, M. R. Hacene, C. Roume, and P. Valtchev. Towards practical tools for mining abstractions in UML models. In *Proc. of the 8th Intl. Conf. on Enterprise Information Systems (ICEIS'06)*, pages 276–283, 2006.
5. S. Ferré, O. Ridoux, and B. Sigonneau. Arbitrary Relations in Formal Concept Analysis and Logical Information Systems. In *Proc. of the 13th Intl. Conf. on Conceptual Structures (ICCS'05)*, volume 3596 of *LNCS*, pages 166–180. Springer, 2005.
6. B. Ganter, G. Stumme, and R. Wille, editors. *Formal Concept Analysis: Foundations and Applications*. LNAI 3626. Springer, 2005.
7. B. Ganter and R. Wille. *Formal Concept Analysis, Mathematical Foundations*. Springer--Verlag, 1999.
8. M. Huchard, M. Hacene, C. Roume, and P. Valtchev. Relational concept discovery in structured datasets. *Annals of Mathematics and Artificial Intelligence*, 49(1):39–76, 2007.
9. S. Kramer, N. Lavrač, and P. Flach. Propositionalization approaches to relational data mining. In S. Džeroski and N. Lavrač, editors, *Relational Data Mining*, pages 262–291. Springer, 2001.
10. S. Kuznetsov. Learning of Simple Conceptual Graphs from Positive and Negative Examples. In *Proc. of the 3rd European Conf. on Principles of KDD (PKDD'99)*, volume 1704 of *LNCS*, pages 384–391. Springer, 1999.
11. J. Lehmann and P. Hitzler. Concept learning in description logics using refinement operators. *Mach. Learn.*, 78:203–250, January 2010.
12. M. Liquière and J. Sallantin. Structural Machine Learning with Galois Lattice and Graphs. In *Proc. of the 15th Intl. Conf. on Machine Learning (ICML'98)*, pages 305–313, 1998.
13. N. Moha, N. Rouane-Hacene, P. Valtchev, and Y.-G. Guéhéneuc. Refactorings of Design Defects using Relational Concept Analysis. In *Proc. of the 6th Intl. Conf. on Formal Concept Analysis (ICFCA'08)*, volume 4933 of *LNCS*, pages 289–304. Springer, 2008.
14. S. Prediger and G. Stumme. Theory-driven logical scaling. In *Proc. 6th Intl. WS Knowledge Representation Meets Databases*, CEUR WS Proc., pages 46–49, 1999.
15. S. Prediger and R. Wille. The Lattice of Concept Graphs of a Relationally Scaled Context. In *Proc. of the 7th Intl. Conf. on Conceptual Structures (ICCS'99)*, pages 401–414. Springer, 1999.
16. M. Rouane-Hacene, M. Huchard, A. Napoli, and P. Valtchev. Relational Concept Analysis: Mining Concept Lattices From Multi-Relational Data (26 p.). *to appear in Annals of Mathematics and Artificial Intelligence*, 2013.
17. S. Rudolph. Exploring Relational Structures via FLE. In *Proc. of the 12th Intl. Conf. on Conceptual Structures (ICCS'04), Huntsville (AL)*, volume 3127 of *LNAI*, pages 196–212. Springer, 2004.
18. P. Valtchev, D. Grosser, C. Roume, and M. Rouane-Hacene. GALICIA: an open platform for lattices. In *Using Conceptual Structures: Contrib. to 11th Intl. Conf. ICCS'03*, pages 241–254. Shaker Verlag, 2003.
19. P. Valtchev, R. Missaoui, and P. Lebrun. A partition-based approach towards building Galois (concept) lattices. *Discrete Mathematics*, 256(3):801–829, 2002.