TECHNISCHE UNIVERSITÄT DRESDEN

Faculty of Computer Science • Institute of Theoretical Computer Science • Chair of Automata Theory

Efficient Axiomatization of OWL 2 EL Ontologies from Data by means of Formal Concept Analysis

Francesco Kriegel

Technische Universität Dresden

37th International Workshop on Description Logics, 18–21 June 2024

Task Description

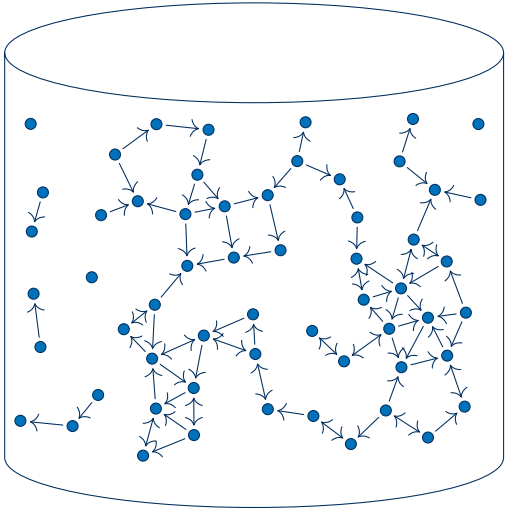
Task Description

Input:

- Graph data with node and edge labels, e.g. RDF triples
- Node labels $x : A$
- Edge labels $(x, y) : r$

Output:

- An OWL 2 EL ontology that is satisfied in the data and that entails all statements satisfied in the data.
- Closure assumption on the data necessary, as otherwise only tautologies hold in the data.



Formal Concept Analysis

Formal Concept Analysis

- Formal Concept Analysis (FCA) is like the description logic $\mathcal{EL}$ without roles and $\bot$.
- Data comes in form of a formal context, which is like an interpretation without roles.
- FCA implications $\{A_1, \ldots, A_m\} \to \{B_1, \ldots, B_n\}$ correspond to DL concept inclusions $A_1 \sqcap \cdots \sqcap A_m \sqsubseteq B_1 \sqcap \cdots \sqcap B_n$.

Definition. An implication base for a formal context $\mathbb{K}$ is an implication set $\mathcal{B}$ that is

- sound: all implications in $\mathcal{B}$ hold in $\mathbb{K}$,
- complete: $\mathcal{B}$ entails all implications that hold in $\mathbb{K}$.

Theorem. Given a formal context $\mathbb{K}$, an implication base can be computed in exponential time and that contains the fewest implications among all bases for $\mathbb{K}$. There are formal contexts that have no polynomial-size base.

Axiomatization of OWL 2 EL Ontologies

No Overfitting in Ontology Learning

If we want to learn or axiomatize an ontology from data that is suitable for real-world applications, then

1. **overfitting must be avoided**: otherwise the input dataset could be simply be rewritten into CIs.
2. **abstraction is necessary**: in order to understand a concept, it is often better to find the commonalities of all objects in this concept instead of just memorizing the single objects and their descriptions.

We therefore impose the following restrictions on the considered DL.

- No nominals.
- No disjunction.
- No negation.

OWL 2 EL

The profile OWL 2 EL is based on the description logic $\mathcal{EL}^{++}$.

- Complex concepts can be built from the concepts A and roles r in the signature: $C ::= \top \mid \bot \mid A \mid C \sqcap C \mid \exists r.C$
- Concept inclusions (CIs) $C \sqsubseteq D$
- Role inclusions (RIs) $R \sqsubseteq s$ where $R ::= \varepsilon \mid r \mid R \circ R$
- Range restrictions (RRs) $\top \sqsubseteq \forall r.C$
- Syntactic sugar: disjointness axioms $C_1 \sqcap \cdots \sqcap C_n \sqsubseteq \bot$, concept equivalences $C \equiv D$, domain restrictions $\exists r.\top \sqsubseteq C$, role equivalences $r \equiv s$, transitivity axioms $r \circ r \sqsubseteq r$, reflexivity axioms $\varepsilon \sqsubseteq r$

Similar to FCA, can we also compute bases of CIs, RIs, and RRs?

A Formal Definition

Definition. Given an interpretation $\mathcal{I}$ and a TBox $\mathcal{T}$ satisfied in $\mathcal{I}$, a base of CIs, RIs, and RRs for $\mathcal{I}$ relative to $\mathcal{T}$ is a TBox $\mathcal{B}$ that is

- sound: all CIs, RIs, and RRs in $\mathcal{B}$ hold in $\mathcal{I}$,
- complete: $\mathcal{B} \cup \mathcal{T}$ entails all CIs, RIs, and RRs that hold in $\mathcal{I}$.

Contrary to implication bases in FCA, existence of bases in DL is not obvious since infinitely many concepts can be constructed from a finite signature.

Concept Inclusions

- Recall that FCA only supports conjunction and that a formal context is like an interpretation without roles.
- To employ FCA, we transform the given interpretation $\mathcal{I}$ into a formal context over an extended signature.
- Additional concept names:
  - bottom concept $\bot$
  - existential restrictions $\exists r.C$, but not all of them.

Concept Inclusions

- Recall that FCA only supports conjunction and that a formal context is like an interpretation without roles.
- To employ FCA, we transform the given interpretation $\mathcal{I}$ into a formal context over an extended signature.
- Additional concept names:
  - bottom concept $\bot$
  - existential restrictions $\exists r.C$, but not all of them.
    Since FCA cannot look into the internal structure of these additional concept names, we require that C is a "closure" of $\mathcal{I}$: if the CI $C \sqsubseteq D$ holds in $\mathcal{I}$, then C is subsumed by D.

    Otherwise, FCA could generate implications $\{A\} \to \{B\}$ and $\{\exists r.A\} \to \{\exists r.B\}$, but FCA does not recognize that the second follows from the first with DL semantics.

    The set of all closures equals the set of all characteristic concepts $X^{\mathcal{I}}$ for subsets X of the domain of $\mathcal{I}$, and is thus finite. $X^{\mathcal{I}}$ is the most specific concept that describes the commonalities of all objects in X.

Concept Inclusions

- The given TBox $\mathcal{T}$ is transformed into a set of FCA implications $\mathcal{L}$ over the extended signature.
- To avoid the computation of DL tautologies, $\mathcal{L}$ further contains all FCA implications $\{C\} \to \{D\}$ where $\emptyset \models C \sqsubseteq D$.
- In the end, we compute the implication base of the formal context relative to $\mathcal{L}$.

Theorem. Given a finite interpretation $\mathcal{I}$ and a TBox $\mathcal{T}$ satisfied in $\mathcal{I}$, a CI base for $\mathcal{I}$ relative to $\mathcal{T}$ can be computed in exponential time and, if all CIs in $\mathcal{T}$ have a particular form, that contains the fewest CIs among all bases.

Role Inclusions and Range Restrictions

- For each role s, the following language is regular: $\{ r_1 \cdots r_n \mid \mathcal{I} \models r_1 \circ \cdots \circ r_n \sqsubseteq s \}$.
- The accepting finite automaton is obtained by viewing $\mathcal{I}$ as a finite automaton and usual automaton operations.
- These automata are rewritten into RIs, using the automata states are additional roles.

- For each role r, we compute the range restriction $\top \sqsubseteq \forall r.C$ where C is the characteristic concept of all r-successors in $\mathcal{I}$.

Theorem. For each finite interpretation $\mathcal{I}$, a complete TBox of concept inclusions, range restrictions, and role inclusions satisfied in $\mathcal{I}$ can be computed in exponential time. There are finite interpretations for which such a TBox cannot be of polynomial size.

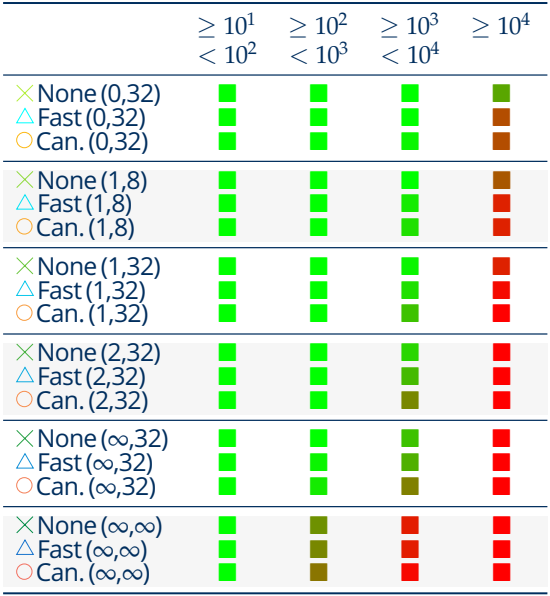
Implementation and Evaluation

Implementation and Evaluation

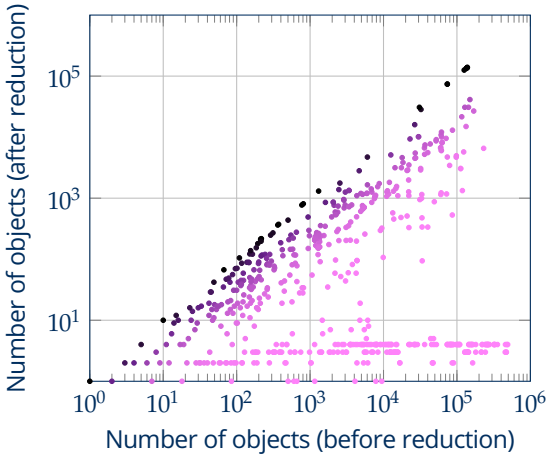
- Prototype in Scala and with Java's Fork/Join framework
- Evaluated on 614 test datasets from real-world domains with up to 747,998 objects
- Computer server: 12 CPU cores at 2.80 GHz and 96 GB main memory (older than 10 years)
- Runtime environment: Oracle GraalVM EE 22.3 (Java 19)
- Resource limits: 8 hours, 80 GB

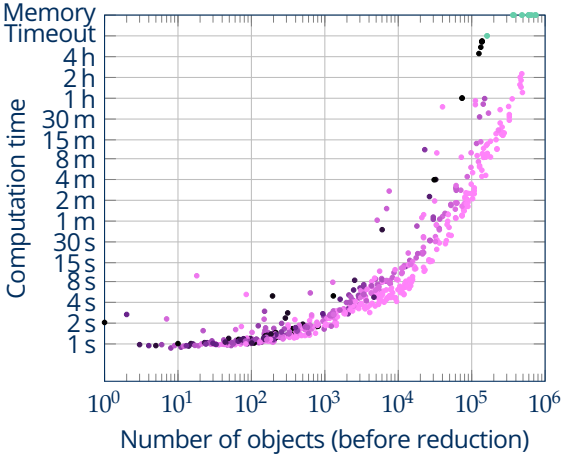
Saving Computation Time:

- Input data is reduced by grouping equi-similar objects (which satisfy the same concepts and concept inclusions).
- Parameter configurations $\mathrm{mode}(\mathrm{rd}, \mathrm{cs})$ where
  - $\mathrm{mode} \in \{\mathrm{Canonical}, \mathrm{Fast}, \mathrm{None}\}$ allows to speed-up or even dispense with computation of disjointness axioms $C_1 \sqcap \cdots \sqcap C_n \sqsubseteq \bot$
  - $\mathrm{rd} \geq 0$ is a role-depth bound or $\mathrm{rd} = \infty$ (unbounded)
  - $\mathrm{cs} \geq 0$ is a conjunction-size limit or $\mathrm{cs} = \infty$ (unlimited)

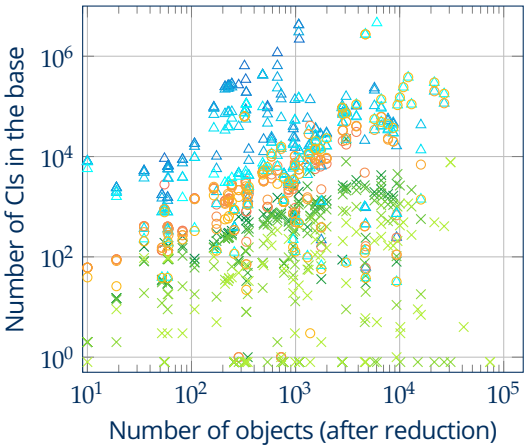

$\geq 10^1$ $< 10^2$
$\geq 10^2$ $< 10^3$
$\geq 10^3$ $< 10^4$
$\geq 10^4$
None (0,32)
Fast (0,32)
Can. (0,32)
None (1,8)
Fast (1,8)
Can. (1,8)
None (1,32)
Fast (1,32)
Can. (1,32)
None (2,32)
Fast (2,32)
Can. (2,32)
None (∞,32)
Fast (∞,32)
Can. (∞,32)
None (∞,∞)
Fast (∞,∞)
Can. (∞,∞)

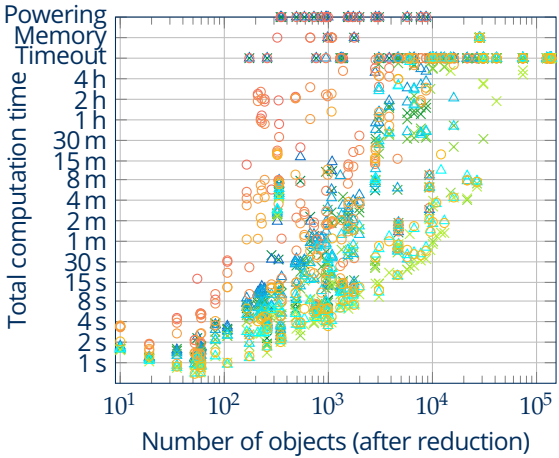
Computing the Dataset Reductions





Computing the Bases of the Reduced Datasets





Do you have questions or comments?