Effective computation of maximal sound approximations of Description Logic ontologies

Marco Console and Jose Mora and Riccardo Rosati and Valerio Santarelli and Domenico Fabio Savo

Dipartimento di Ingegneria Informatica Automatica e Gestionale Antonio Ruberti
Sapienza Università di Roma, Italia

13th International Semantic Web Conference (ISWC-2014)
Riva del Garda, Italy, October 19-23 2014

The problem

- When using ontologies as a formal description of the domain of interest, the use of expressive languages (OWL 2) is useful.
- When using ontologies for reasoning, high expressivity may be a problem. In particular, when accessing large quantities of data (**OBDA**), computational cost of languages such as OWL 2 is prohibitive.

Ontology Approximation: Given a ontology $\mathcal{O}$ in a language $\mathcal{L}$, compute an ontology $\mathcal{O}'$ in a target language $\mathcal{L}'$, in which "as much as possible" of the semantics of $\mathcal{O}$ is preserved.

- ► can represent a solution for performing costly reasoning services over ontologies in expressive languages.

Outline

We investigate the problem of approximating ontologies for OBDA purposes.

1. A new definition of approximation of DL ontologies
2. Approximating OWL 2 ontologies in OWL 2 QL
3. Experimental evaluation
4. Conclusions and future works

Outline

OWL 2 QL is the "data oriented" profile of OWL 2.

Expressions in OWL 2 QL

$$\begin{array}{lcl} B & \longrightarrow & A \mid \exists Q \mid \delta_F(U) \mid \top_C \mid \bot_C \\ C & \longrightarrow & B \mid \neg B \mid \exists Q.A \\ Q & \longrightarrow & P \mid P^- \mid \top_P \mid \bot_P \\ R & \longrightarrow & Q \mid \neg Q \end{array} \qquad \begin{array}{lcl} E & \longrightarrow & \rho(U) \\ F & \longrightarrow & \top_D \mid T_1 \mid \cdots \mid T_n \\ V & \longrightarrow & U \mid \top_A \mid \bot_A \\ W & \longrightarrow & V \mid \neg V \end{array}$$

Assertions in OWL 2 QL

$$B \sqsubseteq C \qquad Q \sqsubseteq R \qquad U \sqsubseteq V \qquad E \sqsubseteq F \qquad ref(P) \qquad irref(P)$$
$$A(a) \qquad P(a,b) \qquad U(a,v)$$

We say that OWL 2 QL is a **closed language**: each set of OWL 2 QL axioms is an OWL 2 QL ontology.

Outline

What we want when we approximate

We deal with *semantic approximation*:

- soundness: only produce correct entailments;
- preserve as much as possible of these entailments by means of an ontology in the target language.

In terms of models of the ontologies:

- Soundness: set of models of the approximation must be a superset of those of the original ontology
- Minimal change: keep minimal distance between the original ontology and its approximation

We give our notion of approximation in a language $\mathcal{L}_T$ of an ontology $\mathcal{O}$.

Definition

Let $\mathcal{O}_S$ be a satisfiable $\mathcal{L}_S$-ontology, and let $\Sigma_{\mathcal{O}_S}$ be the set of predicate and constant symbols occurring in $\mathcal{O}_S$. An $\mathcal{L}_T$-ontology $\mathcal{O}_T$ over $\Sigma_{\mathcal{O}_S}$ is a *global semantic approximation* (**GSA**) in $\mathcal{L}_T$ of $\mathcal{O}_S$ if both the following statements hold:

(i) $Mod(\mathcal{O}_S) \subseteq Mod(\mathcal{O}_T)$; (soundness)

(ii) there is no $\mathcal{L}_T$-ontology $\mathcal{O}'$ over $\Sigma_{\mathcal{O}_S}$ such that $Mod(\mathcal{O}_S) \subseteq Mod(\mathcal{O}') \subset Mod(\mathcal{O}_T)$. (minimal change)

We denote with $globalApx(\mathcal{O}_S, \mathcal{L}_T)$ the set of all the GSAs in $\mathcal{L}_T$ of $\mathcal{O}_S$.

Existence and Uniqueness

Lemma (Existence)

Given a language $\mathcal{L}_T$ and a finite signature Σ, if the set of non-equivalent axioms in an $\mathcal{L}_T$-ontology that one can generate over Σ is finite, then for any $\mathcal{L}_S$-ontology $\mathcal{O}_S$ $globalApx(\mathcal{O}_S, \mathcal{L}_T) \neq \emptyset$.

Lemma (Uniqueness)

Let $\mathcal{L}_T$ be a closed language, and let $\mathcal{O}_S$ be an ontology. For each $\mathcal{O}'$ and $\mathcal{O}''$ belonging to $globalApx(\mathcal{O}_S, \mathcal{L}_T)$, we have that $\mathcal{O}'$ and $\mathcal{O}''$ are logically equivalent.

A constructive notion of approximation

A more constructive definition, based on the notion of *Entailment Set* [Pan&Thomas 2007].

Definition

Let $\mathcal{O}$ be a satisfiable ontology expressed in a language $\mathcal{L}$ over a signature Σ, and let $\mathcal{L}$' be a language, not necessarily different from $\mathcal{L}$. The *entailment set* of $\mathcal{O}$ with respect to $\mathcal{L}'$, denoted as $\mathsf{ES}(\mathcal{O}, \mathcal{L}')$, is the set which contains all $\mathcal{L}'$ axioms over Σ that are entailed by $\mathcal{O}$.

Theorem

Let $\mathcal{O}_S$ be a satisfiable $\mathcal{L}_S$-ontology and let $\mathcal{O}_T$ be a satisfiable $\mathcal{L}_T$-ontology. We have that:

(a) $Mod(\mathcal{O}_S) \subseteq Mod(\mathcal{O}_T)$ if and only if $\mathsf{ES}(\mathcal{O}_T, \mathcal{L}_T) \subseteq \mathsf{ES}(\mathcal{O}_S, \mathcal{L}_T)$;

(b) there is no $\mathcal{L}_T$-ontology $\mathcal{O}'$ such that $Mod(\mathcal{O}_S) \subseteq Mod(\mathcal{O}') \subset Mod(\mathcal{O}_T)$ if and only if there is no $\mathcal{L}_T$-ontology $\mathcal{O}''$ such that $\mathsf{ES}(\mathcal{O}_T, \mathcal{L}_T) \subset \mathsf{ES}(\mathcal{O}'', \mathcal{L}_T) \subseteq \mathsf{ES}(\mathcal{O}_S, \mathcal{L}_T)$.

[Pan&Thomas 2007]: Jeff Z. Pan and Edward Thomas. Approximating OWL-DL ontologies. In Proc. of AAAI, 2007.

Computing the entailment set of an ontology is hard: need to reason over the ontology as a whole!

Idea: instead of reasoning over the whole ontology, we only reason over portions of it.

k-approximation

- **Parametric approximation**: only reason over **k** axioms at a time.
- approximate by computing **GSA** of each set of **k axioms** of the original ontology in isolation.

k-approximation

Definition

Let $\mathcal{O}_S$ be a satisfiable $\mathcal{L}_S$-ontology and let $\Sigma_{\mathcal{O}_S}$ be the set of predicate and constant symbols occurring in $\mathcal{O}_S$. Let $\mathcal{U}_k = \{\mathcal{O}_i^j \mid \mathcal{O}_i^j \in globalApx(\mathcal{O}_i, \mathcal{L}_T),$ such that $\mathcal{O}_i \in subset_k(\mathcal{O}_S)\}$. An $\mathcal{L}_T$-ontology $\mathcal{O}_T$ over $\Sigma_{\mathcal{O}_S}$ is a **k-approximation** in $\mathcal{L}_T$ of $\mathcal{O}_S$ if both the following statements hold:

- $\bigcap_{\mathcal{O}_i^j \in \mathcal{U}_k} Mod(\mathcal{O}_i^j) \subseteq Mod(\mathcal{O}_T)$; (soundness)
- there is no $\mathcal{L}_T$-ontology $\mathcal{O}'$ over $\Sigma_{\mathcal{O}_S}$ such that $\bigcap_{\mathcal{O}_i^j \in \mathcal{U}_k} Mod(\mathcal{O}_i^j) \subseteq Mod(\mathcal{O}') \subset Mod(\mathcal{O}_T)$. (minimal change)

$subset_k(\mathcal{O})$: set of all sets of cardinality k of axioms of $\mathcal{O}$

Two notable cases for k-approximation

- for $k = |\mathcal{O}_S|$, k-approximation = **GSA**

- for $k = 1$, each axiom is treated in isolation, so we consider ontologies formed by a single axiom
  - **Local Semantic Approximation (LSA)**

An example of GSA and LSA

Example

Approximation of $\mathcal{O}$ in OWL 2

$$\mathcal{O} = \{ \begin{array}{lll} A \sqsubseteq B \sqcup C & B \sqsubseteq D & A \sqsubseteq \exists R.D \\ B \sqcap C \sqsubseteq F & C \sqsubseteq D & \exists R.D \sqsubseteq E \end{array} \}.$$

$$\mathcal{O}_{GSA} = \{ \begin{array}{llll} A \sqsubseteq D & B \sqsubseteq D & A \sqsubseteq \exists R & A \sqsubseteq \exists R.D \\ A \sqsubseteq E & C \sqsubseteq D & D \sqsubseteq F \end{array} \}.$$

$$\mathcal{O}_{LSA} = \{ \begin{array}{ll} B \sqsubseteq D & A \sqsubseteq \exists R \\ C \sqsubseteq D & A \sqsubseteq \exists R.D \end{array} \}.$$

Observation: $Mod(\mathcal{O}) \subset Mod(\mathcal{O}_{GSA}) \subset Mod(\mathcal{O}_{LSA})$

- $\mathcal{O}_{GSA}$ approximates $\mathcal{O}$ better than $\mathcal{O}_{LSA}$

Outline

k-approximation in OWL 2 QL

Theorem

Let $\mathcal{O}_S$ be a satisfiable OWL 2 ontology. Then the OWL 2 QL ontology $\bigcup_{\mathcal{O}_i \in subset_k(\mathcal{O}_S)}$ ES$(\mathcal{O}_i, OWL\ 2\ QL)$ is the k-approximation in OWL 2 QL of $\mathcal{O}_S$.

- for OWL 2 QL, the set of non-equivalent axioms that can be generated over a signature is finite $\rightarrow$ **GSA** in OWL 2 QL always exists (Existence Lemma);
- OWL 2 QL is closed:
  - all ontologies in ES$(\mathcal{O}_S, OWL\ 2\ QL)$ are pairwise logically equivalent (Uniqueness Lemma)
  - ES$(\mathcal{O}_S, OWL\ 2\ QL)$ is an OWL 2 QL ontology for any language of $\mathcal{O}_S$
  - the union of a set of OWL 2 QL ontologies is still an OWL 2 QL ontology

Observation: for $k = |\mathcal{O}_S|$ the k-approximation $\mathcal{O}_T$ in OWL 2 QL of $\mathcal{O}_S$ is unique and coincides with its entailment set in OWL 2 QL.

Algorithm: $computeKApx(\mathcal{O}, k)$
Input: a satisfiable OWL 2 ontology $\mathcal{O}$, a positive integer k such that $k \leq |\mathcal{O}|$
Output: an OWL 2 QL ontology $\mathcal{O}_{Apx}$
begin
  $\mathcal{O}_{Apx} \leftarrow \emptyset$;
  foreach ontology $\mathcal{O}_i \in subset_k(\mathcal{O}_S)$
    $\mathcal{O}_{Apx} \leftarrow \mathcal{O}_{Apx} \cup \mathsf{ES}(\mathcal{O}_i, OWL\ 2\ QL)$;
  return $\mathcal{O}_{Apx}$;
end

Outline

Experimental evaluation

1. **GSA** ($k = |\mathcal{O}_S|$) vs. **LSA** ($k = 1$)
2. **GSA** and **LSA** vs. **syntactic** sound approximation (baseline)

- Timeout set at 8 hours
- 41 Bioportal ontologies tested
- GSA computable for 26/41 ontologies
- LSA always computable

Ontology	GSA/original	LSA/GSA	SYNT/GSA	SYNT/LSA	GSA time (s)	LSA time (s)
Vertebrate anatomy	93%	97%	56%	67%	3	3
Spatial	63%	86%	42%	52%	9	4
Translational medicine	86%	99%	30%	64%	19	7
Skeletal anatomy	95%	92%	57%	99%	27	5
Pato	89%	100%	78%	99%	99	18
Lipid	87%	97%	89%	97%	47	10
Plant	96%	81%	81%	100%	929	15
Mosquito anatomy	99%	44%	44%	100%	214	16
Idomal namespace	99%	98%	59%	100%	496	16
Cognitive atlas	97%	100%	26%	30%	162	17
Fly anatomy	99%	67%	67%	100%	25499	45
OVERALL AVERAGE	80%	87%	72%	90%	1110	41

Ontology	LSA/original	LSA time (s)
Protein	47%	20
Dolce	78%	8
Galen-A	70%	26
Fyp	85%	43
Gene	99%	178
FMA OBO	99%	113
OVERALL AVERAGE	72%	51

Final considerations:

- **GSA** provides maximal sound approximation in reasonable time for majority of tested ontologies (80% average for 26/41);
- **LSA** provides very fast solution in all cases, and captures on average significant portion of **GSA** (87% average);
- **LSA** provides good approximation even for ontologies for which **GSA** is not computable (72% average);
- **LSA** and **GSA** both compare favorably against syntactix sound approximation (respectively 90% and 72% average).
  - for very large ontologies, 10% difference for **LSA/SYNT** means preserving thousands of axioms in very little time

GSA and **LSA** both useful approaches!

Outline

Conclusions

- We have proposed a parameterized semantics for computing sound approximations of ontologies;
- We have provided algorithms for approximations (GSA and LSA) of OWL 2 ontologies in OWL 2 QL;
- Extensive experimental evaluation which demonstrate the validity of both GSA and LSA.

Future Works

- Develop techniques for k-approximations with $1 < k < |\mathcal{O}_S|$;
- Integrate ontology module extraction techniques;
- When is LSA enough?
- Generalizing our approach to OBDA scenario: approximating a source ontology with both ontology and mappings in the target OBDA specification.

Thank you!

Entailment Set: huge number of OWL 2 reasoner invocations for axiom implications.

Strategy: exploiting acquired knowledge in order to limit the number of invocation the OWL 2 reasoner (each invocation is $2\textsc{NExpTime}$!).

So...

- if $\exists R$ has no subsumees, then $\exists R.A$, $\exists R.\exists P.A$, $\exists R.\exists P.\exists P^{-}$, ... also have no subsumees.
  - save invocations of the OWL 2 reasoner for subsumees of $\exists R.A$, $\exists R.\exists P.A$, $\exists R.\exists P.\exists P^{-}$, ...
- if $B_1 \sqsubseteq B_2$ and B_1 has no disjoint concepts, then also B_2 has no disjoint concepts.
  - save invocations of the OWL 2 reasoner for disjoint concepts of every subsumer of B_1.