DL-Lite and Conjunctive Queries Extended by Optional Matching (Extended Abstract)*

Shqiponja Ahmetaj, Wolfgang Fischl, Reinhard Pichler, Mantas Šimkus, and Sebastian Skritek

Institute for Information Systems, TU Vienna

1 Introduction

Conjunctive Queries (CQs) constitute the core of most query languages and have been studied intensively in several areas. For querying incomplete data, CQs however suffer one major drawback: they require the complete query to be matched into the data to return answers. One extension of CQs that tries to overcome this problem are *well-designed pattern trees (wdPTs)* [9]. Developed in the context of the Semantic Web, wdPTs are equivalent to a well-behaved fragment of $\{\mathsf{AND}, \mathsf{OPT}\}$-queries of SPARQL [12], and allow a user to retrieve *partial answers* to a query.

Because of this feature, however, wdPTs are nonmonotone. This is problematic for query answering in the presence of implicit knowledge – expressed e.g. by an ontology specified in some Description Logic (DL) – since the usual certain answer semantics turns out to be unsatisfactory in this setting. We observe that the recently released recommendation of the SPARQL entailment regimes [6] provides a semantics exactly for this case. However, it is defined in a simpler and less expressive way than the certain answers semantics, and does not utilize the full potential of the implicit information.

The **goal of this work** is to introduce an intuitive certain answer semantics for the class of well-designed pattern trees under DL-Lite$_\mathcal{R}$ (which provides the theoretical underpinning of the OWL 2 QL entailment regime). After introducing wdPTs, we first discuss some of the problems with an adoption of a certain answer semantics for them and propose a suitable modified definition. We then briefly present results on the complexity of typical reasoning tasks.

Related Work to our findings includes the work our approaches are based upon [3–6]. There is a huge body of results on CQ answering under different DLs (cf. [4, 5, 11, 13]). For SPARQL recent work [8] presents a *stronger* semantics, where entire mappings are discarded, whose possible extensions to optional subqueries would imply inconsistencies in the knowledge base. Further related work includes [2, 7, 10] which is discussed in the long version of this paper.

* A longer version of this paper has been accepted for publication at WWW 2015 [1].

2 DL-Lite$_\mathcal{R}$ and Well-designed Pattern Trees

We assume the reader to be familiar with DL-Lite$_\mathcal{R}$ [4]. A DL-Lite$_\mathcal{R}$ *knowledge base* (KB) is a tuple $\mathcal{K} = \langle \mathcal{A}, \mathcal{T} \rangle$, where $\mathcal{A}$ is an *ABox* and $\mathcal{T}$ is a *TBox*. The definition of an interpretation $\mathcal{I} = (\Delta^\mathcal{I}, \cdot^\mathcal{I})$ is the usual one. In addition, we make the *standard name assumption (SNA)*, i.e. we assume that $\Delta^\mathcal{I}$ contains all individuals, and that $a^\mathcal{I} = a$ for each individual a.

A *well-designed pattern tree* $\mathcal{P}$ is a tuple $(T, \lambda, \boldsymbol{x})$ such that:

1. T is a rooted tree and λ maps each node t in T to a *conjunctive query (CQ)*. A CQ here is a set of *atoms*, where atoms are built as usual, i.e. from concept and role names together with individuals and variables.
2. For every variable y occurring in T, the set of nodes containing y is connected.
3. $\boldsymbol{x}$ is a tuple of variables from T, called the *free variables* of $\mathcal{P}$.

Intuitively, the parent-child relationships in the tree express *optional matching*. I.e., the result of the "parent-CQ" shall be extended by the "child-CQ" if possible — otherwise the child shall be ignored, and only the result of the parent is returned. Finally $\boldsymbol{x}$ are the "output" variables.

A *mapping* μ is any partial function whose domain $\mathsf{dom}(\mu)$ contains only variables. We say a mapping μ_1 is subsumbed by another mapping μ_2, denoted by $\mu_1 \sqsubseteq \mu_2$, if $\mathsf{dom}(\mu_1) \subseteq \mathsf{dom}(\mu_2)$ and $\mu_1(x) = \mu_2(x)$ for all $x \in \mathsf{dom}(\mu_1)$. Also, for a mapping μ and some property A, we shall say that μ is $\sqsubseteq$*-maximal w.r.t.* A if μ satisfies A, and there is no μ' such that $\mu \sqsubseteq \mu'$, $\mu' \not\sqsubseteq \mu$, and μ' satisfies A. For any mapping μ and a tuple of variables $\boldsymbol{x}$, we denote by $\mu_{\boldsymbol{x}}$ the restriction of μ to the variables in $\boldsymbol{x}$.

The notion of a mapping μ that is a *match* for a CQ q in an interpretation $\mathcal{I}$ is defined in the standard way. Assume a wdPT $\mathcal{P} = (T, \lambda, \boldsymbol{x})$ and an interpretation $\mathcal{I}$. For an initial segment T' of T, i.e. a connected subgraph containing the root of T, we define $q_{T'}$ to be the CQ $\bigcup_{t \in T'} \lambda(t)$. Then a *match* for $\mathcal{P}$ in $\mathcal{I}$ is any mapping μ such that μ is a match for $q_{T'}$ in $\mathcal{I}$ for some initial segment T' of T. Let M be the set of all $\sqsubseteq$-maximal matches from $\mathcal{P}$ to $\mathcal{I}$. Then the result of evaluating $\mathcal{P}$ over $\mathcal{I}$, projected to $\boldsymbol{x}$, is the set $[\![\mathcal{P}]\!]_\mathcal{I} = \{\mu_{\boldsymbol{x}} \mid \mu \in M\}$. Note that the order of child nodes in such tress does not affect the query answer (see [9, 12]).

In the following example, we illustrate wdPTs as well as the reason why the usual certain answer semantics (i.e., a tuple is a certain answer if it is present in every model) turns out to be unsatisfactory in our setting:

Example 1. Let $\mathcal{P}$ be the wdPT $(T, \lambda, \boldsymbol{x})$ where T consists of the root r with the single child t, $\lambda(r) = \{\mathsf{teaches}(x, y)\}$, $\lambda(t) = \{\mathsf{knows}(y, z)\}$, and $\boldsymbol{x} = \{x, z\}$. Consider the KB $\mathcal{K}$ consisting of an ABox $\mathcal{A} = \{\mathsf{Prof}(b)\}$, and a TBox $\mathcal{T} = \{(\mathsf{Prof} \sqsubseteq \exists \mathsf{teaches})\}$. Let $\mathcal{I}$ be as follows: $\mathsf{Prof}^\mathcal{I} = \{(b)\}$. Clearly, $\mathcal{I} \models \mathcal{K}$. The query yields in $\mathcal{I}$ as only answer the mapping $\mu = \{x \to b\}$. Clearly, also the interpretation $\mathcal{I}'$, where $\mathsf{Prof}^{\mathcal{I}'} = \{(b)\}$, $\mathsf{teaches}^{\mathcal{I}'} = \{(b, c)\}$ and $\mathsf{knows}^{\mathcal{I}'} = \{(c, d)\}$ is a model of $\mathcal{K}$. But in $\mathcal{I}'$, μ is no longer an answer since μ can be extended to answer $\mu' = \{x \to b, z \to d\}$. Hence, there is no mapping which is an answer in every possible model of $\mathcal{K}$. □

Definition 1. *Let $\mathcal{K} = (\mathcal{A}, \mathcal{T})$ be a KB and $\mathcal{P} = (T, \lambda, \boldsymbol{x})$ a wdPT. A mapping μ is a* certain answer *to $\mathcal{P}$ over $\mathcal{K}$ if it is a $\sqsubseteq$-maximal mapping s.t. (1) $\mu \sqsubseteq [\![\mathcal{P}]\!]_{\mathcal{I}}$ for every model $\mathcal{I}$ of $\mathcal{K}$, and (2) $\mathsf{vars}(q_{T'}) \cap \boldsymbol{x} = \mathsf{dom}(\mu)$ for some initial segment T' of T. We denote by $\mathsf{cert}(\mathcal{P}, \mathcal{K})$ the set of certain answers to $\mathcal{P}$ over $\mathcal{K}$.*

The reason for restricting the set of certain answers to $\sqsubseteq$-maximal mappings is that wdPTs in general may have "subsumed" answers, i.e. mappings s.t. also some proper extension is an answer. But then – with set semantics – we cannot recognize the reason why some subsumed answer is possibly not an answer in some possible world. Therefore, in our first step towards extending CQs by optional matching, we allow only "maximal" answers as certain answers.

Property (2) ensures that the domain of such an answer adheres to the tree structure of the wdPT. However, we can show that this can be enforced in a simple post-processing step. Likewise, also projection can be deferred to a post-processing step. The task is thus to compute a set $\mathsf{certp}(\mathcal{P}, \mathcal{K})$ of *certain pre-answers* (i.e., mappings that satisfy Definition 1 except property (2), ignoring projection), which can be done via the *canonical model*. For a given KB $\mathcal{K}$, we assume a canonical model of $\mathcal{K}$, denoted as $\mathsf{can}(\mathcal{K})$, to be defined as in [4].

Theorem 1. *Let $\mathcal{K} = (\mathcal{A}, \mathcal{T})$ be a KB and $\mathcal{P}$ a wdPT. Then, $\mathsf{certp}(\mathcal{P}, \mathcal{K}) = \mathrm{MAX}([\![\mathcal{P}]\!]_{\mathsf{can}(\mathcal{K})}\downarrow)$, where $\mathrm{MAX}(M)$ is the set of $\sqsubseteq$-maximal mappings in M, $M\downarrow := \{\mu\downarrow \mid \mu \in M\}$ ($\mu\downarrow$ is the restriction of μ to those variables which are mapped to the individual names that occur in $\mathcal{A}$).*

To cope with the potential infinite canonical model, query rewriting algorithms have been developed in the literature. By introducing several adaptations and extensions of the rewriting-based CQ evaluation for DL-Lite from [4], we develop two different algorithms to answer wdPTs over DL-Lite$_{\mathcal{R}}$ KBs. [1] Based on these rewriting algorithms, we analyze the complexity of query answering and of several static query analysis tasks such as query containment and equivalence. We are able to show that the additional power of our new semantics comes without additional costs in terms of complexity.

For future work, we want to investigate further more expressive DLs under our certain answer semantics. The implementation of the rewriting algorithms and the development of suitable benchmarks, is a challenging task as well. Additionally, we will extend our work to allow TBox queries and other fragments of SPARQL.

Acknowledgements

This work was supported by the Vienna Science and Technology Fund (WWTF), project ICT12-15 and by the Austrian Science Fund (FWF): P25207-N23 and W1255-N23.

[1] Note that, in the full version we consider a fragment of well-designed SPARQL under OWL 2 QL entailment, which corresponds exactly to what we consider here.

References

1. S. Ahmetaj, W. Fischl, R. Pichler, M. Šimkus, and S. Skritek. Towards reconciling SPARQL and certain answers, 2014. Accepted for publication, WWW 2015.
2. M. Arenas, G. Gottlob, and A. Pieris. Expressive languages for querying the semantic web. In *Proc. of PODS 2014*, pages 14–26. ACM, 2014.
3. M. Arenas and J. Pérez. Querying semantic web data with SPARQL. In *Proc. of PODS 2011*, pages 305–316. ACM, 2011.
4. D. Calvanese, G. De Giacomo, D. Lembo, M. Lenzerini, and R. Rosati. Tractable reasoning and efficient query answering in description logics: The DL-Lite family. *J. Autom. Reasoning*, 39(3):385–429, 2007.
5. T. Eiter, M. Ortiz, M. Šimkus, T. Tran, and G. Xiao. Query rewriting for Horn-$\mathcal{SHIQ}$ plus rules. In *Proc. of AAAI 2012*. AAAI Press, 2012.
6. B. Glimm and C. Ogbuji. SPARQL 1.1 Entailment Regimes. W3C Recommendation, W3C, Mar. 2013. `http://www.w3.org/TR/sparql11-entailment`.
7. R. Kontchakov, M. Rezk, M. Rodriguez-Muro, G. Xiao, and M. Zakharyaschev. Answering SPARQL queries over databases under OWL 2 QL entailment regime. In *Proc. of ISWC 2014*, pages 552–567. Springer, 2014.
8. E. V. Kostylev and B. C. Grau. On the semantics of SPARQL queries with optional matching under entailment regimes. In *Proc. of ISWC 2014*, pages 374–389. Springer, 2014.
9. A. Letelier, J. Pérez, R. Pichler, and S. Skritek. Static analysis and optimization of semantic web queries. *ACM Trans. Database Syst.*, 38(4):25, 2013.
10. L. Libkin. Incomplete data: what went wrong, and how to fix it. In *Proc. PODS 2014*, pages 1–13. ACM, 2014.
11. M. Ortiz, D. Calvanese, and T. Eiter. Data complexity of query answering in expressive description logics via tableaux. *Journal of Automated Reasoning*, 41(1):61–98, 2008.
12. J. Pérez, M. Arenas, and C. Gutierrez. Semantics and complexity of SPARQL. *ACM Trans. Database Syst.*, 34(3), 2009.
13. R. Rosati. On conjunctive query answering in EL. In *Proc. DL 2007*. CEUR-WS.org, 2007.