Relaxing and Restraining Queries for OBDA (Extended Abstract)*

Medina Andreşel and **Yazmín Ibáñez-García** and **Magdalena Ortiz** and **Mantas Šimkus**
{andresel,ibanez,ortiz}@kr.tuwien.ac.at | simkus@dbai.tuwien.ac.at
Faculty of Informatics, TU Wien, Austria

Abstract

We investigate query reformulation rules in OBDA to obtain either more or less answers. We extend *DL-Lite* with complex role inclusions and define rules that produce query relaxations/restrictions over any dataset. We also introduce a set of data-driven rules to get more fine-grained reformulations.

In *Ontology-based data access (OBDA)* an ontology provides a conceptual view of a collection of data sources, and describes knowledge about the domain of interest at a high level of abstraction. Thus users can formulate queries over data sources using a familiar vocabulary provided by the ontology, while the represented knowledge can be leveraged to retrieve more complete answers. For example, consider the following dataset about cultural events and their locations,

$$\begin{aligned}\mathcal{A}_e = \{&\mathrm{Concert}(\mathsf{c_1}), \mathrm{Venue}(\mathsf{StateOpera}), \mathrm{Exhibition}(\mathsf{ex_1}),\\ &\mathrm{City}(\mathsf{Vienna}), \mathrm{CulturEvent}(\mathsf{ev_1}), \mathrm{Country}(\mathsf{Austria}),\\ &\mathrm{occursIn}(\mathsf{c_1}, \mathsf{StateOpera}), \mathrm{occursIn}(\mathsf{ex_1}, \mathsf{Vienna}),\\ &\mathrm{occursIn}(\mathsf{ev_1}, \mathsf{Austria}), \mathrm{locIn}(\mathsf{StateOpera}, \mathsf{Vienna}),\\ &\mathrm{locIn}(\mathsf{Vienna}, \mathsf{Austria})\},\end{aligned}$$

and the ontology $\mathcal{T}_e$ below, which captures, among other things, the knowledge that both concerts and exhibitions are cultural events.

$$\begin{aligned}\mathrm{Country} &\sqsubseteq \mathrm{Location} & \mathrm{Exhibition} &\sqsubseteq \mathrm{CulturEvent}\\ \mathrm{Venue} &\sqsubseteq \mathrm{Location} & \mathrm{Theater} &\sqsubseteq \mathrm{Venue}\\ \mathrm{City} &\sqsubseteq \mathrm{Location} & \mathrm{Museum} &\sqsubseteq \mathrm{Venue}\\ \mathrm{Concert} &\sqsubseteq \mathrm{CulturEvent} & \mathrm{CulturEvent} &\sqsubseteq \mathrm{Event}\\ \exists \mathrm{locIn} &\sqsubseteq \mathrm{Location} & \exists \mathrm{locIn}^- &\sqsubseteq \mathrm{Location}\\ \exists \mathrm{occursIn} &\sqsubseteq \mathrm{Event} & \exists \mathrm{occursIn}^- &\sqsubseteq \mathrm{Location}\end{aligned}$$

Using this knowledge one can retrieve all cultural events, $\mathsf{ex_1}$, $\mathsf{ev_1}$, and $\mathsf{c_1}$, by posing the query:

$$q_1(x) \leftarrow \mathrm{CulturEvent}(x).$$

Description logics (DLs) of the *DL-Lite* family have been particularly tailored for OBDA (Calvanese et al. 2007). As a result, queries mediated by *DL-Lite* ontologies are first-order (FO)-rewritable. This means that evaluating a query q over $(\mathcal{T}, \mathcal{A})$ can be reduced to evaluate a query $q_{\mathcal{T}}$ (incorporating knowledge from $\mathcal{T}$) over $\mathcal{A}$ alone, which amounts to query evaluation in relational databases. In our example, a rewriting of q_1 is

$$q_{\mathcal{T}}(x) \leftarrow \mathrm{CulturEvent}(x) \vee \mathrm{Exhibition}(x) \vee \mathrm{Concert}(x)$$

In this paper we investigate the use of rewritings in OBDA for *relaxing* and *restraining* queries to, respectively, retrieve more or less answers[1]. A key observation in our approach is that query *restrictions* can be obtained using existing rewriting rules and that 'counterparts' of these rules can be defined to produce *relaxations*. In our example, the query $q_c(x) \leftarrow \mathrm{Concert}(x)$ that restricts q_1 occurs as a disjunct in $q_{\mathcal{T}}$.

Using the perfect reformulation for *DL-Lite* proposed by (Calvanese et al. 2007), a rewriting q' of q is obtained by applying an *atom substitution* θ to q as described in rules **(S1)** – **(S5)** in Table 1. A way to define a counterpart e.g., for **(S3)** to obtain a relaxation of a given query q w.r.t. a TBox $\mathcal{T}$ is the rule **(G2)**: replace $A(x) \in q$ by $r(x, y)$ if $A \sqsubseteq \exists r \in \mathcal{T}$, with y a fresh variable. We also define counterparts of rules **S1**–**S5** to obtain query relaxations (see **G1**–**G5** in the full version of this paper).

Notably, there are intuitive answers and reformulations that cannot be produced with the standard *DL-Lite* rewriting rules and their counterparts. For example, consider a query retrieving concerts occurring in Vienna:

$$q_2(x) \leftarrow \mathrm{Concert}(x), \mathrm{occursIn}(x, y), y = \mathsf{Vienna}.$$

There are no answers to q_2 when evaluated over $(\mathcal{T}_e, \mathcal{A}_e)$, although $\mathsf{c_1}$ may be considered an answer to q_2 according to

(S1) if $A_1 \sqsubseteq A_2 \in \mathcal{T}$ and $A_2(x) \in q$, then $\theta = [A_2(x)/A_1(x)]$;

(S2) if $A \sqsubseteq \exists r \in \mathcal{T}$ and $r(x, y) \in q$, and y is a non-answer variable occurring only once in q, then $\theta = [r(x, y)/A(x)]$;

(S3) if $\exists r \sqsubseteq A \in \mathcal{T}$ and $A(x) \in q$, then $\theta = [A(x)/r(x, z^q)]$;

(S4) if $r \sqsubseteq s \in \mathcal{T}$ and $s(x, y) \in q$, then $\theta = [s(x, y)/r(x, y)]$;

(S5) if $r \sqsubseteq s^- \in \mathcal{T}$ and $s(x, y) \in q$, then $\theta = [s(x, y)/r(y, x)]$;

(S6) if $t \cdot s \sqsubseteq r \in \mathcal{T}$ and $r(x, y) \in q$, then $\theta = [r(x, y)/\{t(x, z^q), s(z^q, y)\}]$.

where z^q is a fresh variable.

Table 1: Rewriting rules for *DL-Lite*$^{\mathcal{HR}}$

*This research was funded by FWF Projects P30360 and W1255-N23

[1]The full version of this paper can be found at `https://arxiv.org/pdf/1808.02850.pdf`

the intuition that *if an event occurs in a venue located in a city, then it occurs in that city*. In order to enable this kind of reformulations, we extend the expressive power of *DL-Lite* with *complex role inclusions* (CRIs). In our example, we could add the following:

$$\mathrm{occursIn} \cdot \mathrm{locIn} \sqsubseteq \mathrm{occursIn} \quad (1)$$

to capture the intuition above. We call the extension of *DL-Lite* with CRIs *DL-Lite*$^{\mathcal{HR}}$, and propose reformulation rules operating not only along the subclass (subrole) relation, but also along CRIs (see **(S6)** in Table 1). We can now use (1) and **(S6)** to restrain the query

$$q_3(x) \leftarrow \mathrm{Concert}(x), \mathrm{occursIn}(x, y), \mathrm{City}(y),$$

from *all concerts occurring in a city*, to only those for which a more specific location within a city is known:

$$q_3'(x) \leftarrow \mathrm{Concert}(x), \mathrm{occursIn}(x, z), \mathrm{locIn}(z, y), \mathrm{City}(y)$$

The following rule is the dual of **(S6)**, namely **(G6)** replaces $\{r(x, y), s(y, z)\} \subseteq q$ by $r(x, z)$, if y is a *non-answer variable* that does not occur elsewhere in q and $r \cdot s \sqsubseteq r \in \mathcal{T}$. We show that our ontology-driven rules restrain (or relax) a query q into q' in the following sense: for every dataset $\mathcal{A}$ certain answers of q' w.r.t. $(\mathcal{T}, \mathcal{A})$ are necessarily contained in (or contain) the certain answers of q.

It is well-known that unrestricted usage of CRI can easily lead to undecidability (Horrocks and Sattler 2004). Even when imposing syntactic restrictions to CRIs to regain decidability, such as regularity (Kazakov 2010), query answering is not FO-rewritable in *DL-Lite*$^{\mathcal{HR}}$ (Artale et al. 2009). In the full version of this paper, we propose suitable restrictions to CRIs to regain FO-rewritability.

We also consider *data-driven* reformulations, which unlike the ontology-based ones, are specific to a dataset at hand. For example, the query q_2 asking for *concerts occurring in Vienna* could be restrained to *concerts in the State Opera in Vienna*, or relaxed to *all concerts in Austria*. These reformulations cannot be done on the basis of the TBox alone since they depend on the specific dataset. In our example, they are based on the assertions $\mathrm{locIn}(\mathsf{Vienna}, \mathsf{Austria})$ and $\mathrm{locIn}(\mathsf{StateOper}, \mathsf{Vienna})$.

Our data-driven rules use assertions as follows for $\mathcal{K} = (\mathcal{T}, \mathcal{A})$: **(SD1)** if $A(x) \in q$ and $\mathcal{K} \models A(a)$, then we can restrain q by adding the atom $x = a$; and **(SD2)** if $r(x, y) \in q$ and $\mathcal{K} \models r(a, b)$, then we can add either $x = a$ or $y = b$. For relaxing, **(GD1)** replaces $x = a$ with $A(x)$, if $\mathcal{K} \models A(a)$; and **(GD2)** replaces $x = a \in q$ by the atoms $r(x, y), y = b$, if $\mathcal{K} \models r(a, b)$. For example, using **(GD2)** and $\mathrm{locIn}(\mathsf{Vienna}, \mathsf{Austria}) \in \mathcal{A}_e$, we can relax the query from the *concerts in Vienna*,

$$q(x) \leftarrow \mathrm{Concert}(x), \mathrm{occursIn}(x, y), y = \mathsf{Vienna}$$

to *concerts that occur in a location in Austria*:

$$q'(x) \leftarrow \mathrm{Concert}(x), \mathrm{occursIn}(x, y), \mathrm{locIn}(y, z), z = \mathsf{Austria}.$$

Besides concepts and role assertions, we consider as well *dependencies* that are not necessarily implied by the ontology, but that can be guaranteed to hold in the current dataset. A quick inspection at $\mathcal{A}_e$ reveals that *every existing venue is located in a city*. We could use this knowledge to restrain the query

$$q(x) \leftarrow \mathrm{Event}(x), \mathrm{occursIn}(x, y), \mathrm{locIn}(y, z), \mathrm{City}(z) \quad \text{into}$$

$$q'(x) \leftarrow \mathrm{Event}(x), \mathrm{occursIn}(x, y), \mathrm{Venue}(y).$$

Indeed, such a reformulation could be done based on the ontology alone, provided that the axiom $\mathrm{Venue} \sqsubseteq \exists \mathrm{locIn}.\mathrm{City}$ is present. However, we may not have such an axiom, and it may not be possible or desirable to add it. Then, for a given $\mathcal{K} = (\mathcal{T}, \mathcal{A})$, we define rules using *query containment tests*, $q_1(x) \subseteq_{\mathcal{K}} q_2(x)$, where q_1, q_2 are queries with at most two atoms and two variables. We remark these tests are not expensive in the extension of *DL-Lite* with CRI that is FO-rewritable. These rules are very similar to the ones using only the ontology, except they allow to replace $B(x)$ by $A(x)$ for restraining, not only when $A \sqsubseteq B$ is in $\mathcal{T}$, but also when the weaker condition $A(x) \subseteq_{\mathcal{K}} B(x)$ holds. Note that these replacements are also allowed for some more complex pairs of atoms. For instance, if $q^*(x) \leftarrow r(x, y), B(y)$ and $A(x) \subseteq_{\mathcal{K}} q^*(x)$, rule **(GD4)** will replace $A(x) \in q$ with $r(x, y), B(y)$ producing a relaxation. Consider for example the query

$$q(x) \leftarrow \mathrm{Event}(x), \mathrm{occursIn}(x, y), \mathrm{City}(y)$$

if $\mathrm{City}(x) \subseteq_{\mathcal{K}} \exists \mathrm{locIn}.\mathrm{Country}(x)$, then using rule **(GD4)** we obtain:

$$q'(x) \leftarrow \mathrm{Event}(x), \mathrm{occursIn}(x, y), \mathrm{locIn}(y, z), \mathrm{Country}(z)$$

Now, we can actually apply **(G6)** obtaining the query

$$q'(x) \leftarrow \mathrm{Event}(x), \mathrm{occursIn}(x, z), \mathrm{Country}(z).$$

Thus, our data-driven rules are useful for query reformulation not only on their own, but also because they may trigger other relevant reformulations that were not obtainable otherwise. For the data-driven rules the containment only holds when evaluated over $(\mathcal{T}, \mathcal{A})$, but not for an arbitrary $\mathcal{A}$.

The proposed reformulations can aid users to explore heterogeneous, unstructured and incomplete datasets in the same spirit as online analytical processing (OLAP) supports the exploration of structured data from multiple perspectives and at various granularity levels (Codd, Codd, and Salley 1993). For that purpose, our extension of *DL-Lite* takes into account *dimensional* knowledge, analogous to the so-called *multidimensional data model* (Hurtado and Mendelzon 2002), and our reformulation rules are designed in such a way that they emulate so-called 'rolling-up' and 'drilling down' operations along dimensions.

References

Artale, A.; Calvanese, D.; Kontchakov, R.; and Zakharyaschev, M. 2009. The dl-lite family and relations. *J. Artif. Int. Res.* 36(1):1–69.

Calvanese, D.; De Giacomo, G.; Lembo, D.; Lenzerini, M.; and Rosati, R. 2007. Tractable reasoning and efficient query answering in description logics: The *DL-Lite* family. *J. Autom. Reasoning* 39(3):385–429.

Codd, E. F.; Codd, S. B.; and Salley, C. T. 1993. Providing OLAP to User-Analysts: An IT mandate.

Horrocks, I., and Sattler, U. 2004. Decidability of SHIQ with complex role inclusion axioms. *Artif. Intell.* 160(1-2):79–104.

Hurtado, C. A., and Mendelzon, A. O. 2002. OLAP dimension constraints. In *PODS*, 169–179. ACM.

Kazakov, Y. 2010. An extension of complex role inclusion axioms in the description logic SROIQ. In *Proc. of IJCAR*.