The Combined Approach to Query Answering in *DL-Lite*

Roman Kontchakov

Department of Computer Science and Inf. Systems, Birkbeck College, London

http://www.dcs.bbk.ac.uk/~roman

joint work with

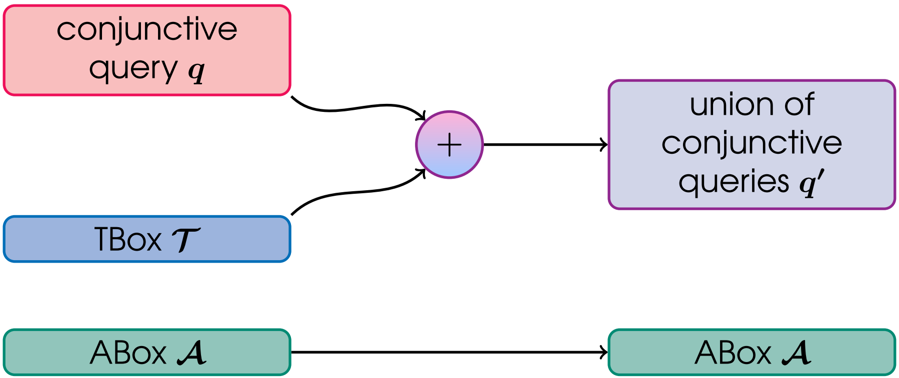
Carsten Lutz, David Toman, Frank Wolter and Michael Zakharyaschev

Pure Query Rewriting

'*DL-Lite* family includes the first DLs specifically tailored for **effective query answering** over large amounts of instances.'

D. Calvanese *et al.*, 2007

effective = in AC^0 for data complexity



Pure Query Rewriting: an Example of PerfectRef

$q(x) \leftarrow \text{TeachesTo}(x, y), \text{HasTutor}(y, z)$

Student $\sqsubseteq$ $\exists$HasTutor $\quad$ $\text{HasTutor}(x_1, y_1) \leftarrow \text{Student}(x_1)$

$q(x) \leftarrow \text{TeachesTo}(x, y), \text{Student}(y)$

$\exists\text{TeachesTo}^- \sqsubseteq$ Student $\quad$ $\text{Student}(x_2) \leftarrow \text{TeachesTo}(y_2, x_2)$

$q(x) \leftarrow \text{TeachesTo}(x, y), \text{TeachesTo}(x_2, y)$

unification

$q(x) \leftarrow \text{TeachesTo}(x, y)$

Professor $\sqsubseteq$ $\exists$TeachesTo $\quad$ $\text{TeachesTo}(x_3, y_3) \leftarrow \text{Professor}(x_3)$

$q(x) \leftarrow \text{Professor}(x)$

$\exists\text{HasTutor}^- \sqsubseteq$ Professor $\quad$ $\text{Professor}(x_4) \leftarrow \text{HasTutor}(y_4, x_4)$

$q(x) \leftarrow \text{HasTutor}(y_4, x)$

Intuitive!

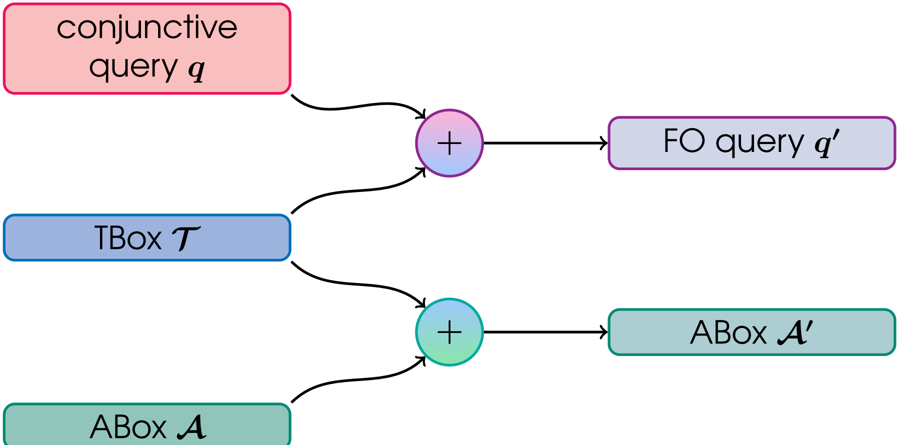
NB. what if Student has many subclasses? $\exists$TeachesTo?

$\mathcal{O}((|\mathcal{T}| \cdot |q|)^{|q|})$ subqueries

Combined Approach in $\mathcal{EL}$

(Lutz, Toman & Wolter, 2008)

query answering in $\mathcal{EL}$ is **PTIME**-complete for data complexity



$\mathcal{A}'$ is computed in polytime in $\mathcal{A}$
and only when $\mathcal{A}$ is updated

Variants of *DL-Lite*

$$R \quad ::= \quad P \quad | \quad P^-$$

$$C \quad ::= \quad \bot \quad | \quad A \quad | \quad \geq k R$$

TBox concept inclusions

DL-Lite$^{\mathcal{N}}_{horn}$: $\quad C_1 \sqcap \dots \sqcap C_n \sqsubseteq C$

DL-Lite$^{\mathcal{N}}_{core}$: $\quad C_1 \sqsubseteq C_2, \quad C_1 \sqsubseteq \neg C_2$

ABox assertions: $\quad C(a), \quad R(a,b)$

DL-Lite$^{\mathcal{F}}_{\alpha}$ $=$ *DL-Lite*$^{\mathcal{N}}_{\alpha}$ with $\exists R$ and $\geq 2\, R \sqsubseteq \bot$ only

DL-Lite$^{(\mathcal{HN})}_{\alpha}$ $=$ *DL-Lite*$^{\mathcal{N}}_{\alpha}$ with (restricted) role inclusions, role disjointness, etc.

DL-Lite$^{(\mathcal{HF})}_{\alpha}$ $=$ *DL-Lite*$^{\mathcal{F}}_{\alpha}$ with (restricted) role inclusions, role disjointness, etc.

In all these languages, answering positive existential queries (under UNA) is in AC^0 for data complexity

positive existential formulas are built from $A(x)$ and $R(x,y)$ using $\exists$, $\wedge$ and $\vee$

ABox Expansion in *DL-Lite*

canonical interpretation $\mathcal{I}_\mathcal{K}$:

$\Delta^\mathcal{I} = \mathsf{Ind}(\mathcal{A}) \cup \{c_R \mid R \text{ is generating in } \mathcal{K}\}$

$a \rightsquigarrow c_{R_1} \rightsquigarrow \cdots \rightsquigarrow c_{R_n}$ $\quad R_n$ is generating

$\mathcal{K} \models \exists R_1(a)$ but $R_1(a,b) \notin \mathcal{A}$ for all $b \in \mathsf{Ind}(\mathcal{A})$

$\mathcal{T} \models \exists R_i^- \sqsubseteq \exists R_{i+1}$ and $R_i^- \neq R_{i+1}$

$A^{\mathcal{I}_\mathcal{K}} = \{a \mid \mathcal{K} \models A(a)\} \cup \{c_R \mid \mathcal{T} \models \exists R^- \sqsubseteq A\}$

$P^{\mathcal{I}_\mathcal{K}} = \{(a,b) \mid P(a,b) \in \mathcal{A}\} \cup \{(d, c_P) \mid d \rightsquigarrow c_P\} \cup \{(c_{P^-}, d) \mid d \rightsquigarrow c_{P^-}\}$

$\mathcal{I}_\mathcal{K}$ is not a model $\qquad \mathcal{T} = \{A \sqsubseteq \exists P, \geq 2\, P^- \sqsubseteq \bot\},\ \ \mathcal{A} = \{A(a), A(b)\}$

$\mathcal{I}_\mathcal{K}$ does not give the right answers

$q = \exists v\, P(v,v), \quad \mathcal{T} = \{A \sqsubseteq \exists P, \exists P^- \sqsubseteq \exists P\}, \quad \mathcal{A} = \{A(a)\}$

$q = \exists v_2\, (P(v_1, v_2) \wedge P(v_3, v_2)), \quad \mathcal{T} = \{A \sqsubseteq \exists P\}, \quad \mathcal{A} = \{A(a), A(b)\}$

The unravelling $\mathcal{U}_\mathcal{K}$ is **almost** a (canonical) model of $\mathcal{I}_\mathcal{K}$ and does give the right answers

Query Rewriting for *DL-Lite*$^{\mathcal{N}}_{horn}$ (1)

we rewrite a given CQ $\boldsymbol{q}$ into an FO query $\boldsymbol{q}^{\dagger}$ such that

- answers to $\boldsymbol{q}$ in $\mathcal{U}_{\mathcal{K}}$ = answers to $\boldsymbol{q}^{\dagger}$ in $\mathcal{I}_{\mathcal{K}}$
- $|\boldsymbol{q}^{\dagger}| = \mathcal{O}(|\boldsymbol{q}| \cdot |\mathcal{T}|)$

$$\boldsymbol{q}^{\dagger} = \exists \vec{\boldsymbol{u}}\, (\varphi \wedge \varphi_1 \wedge \varphi_2 \wedge \varphi_3)$$

$$\varphi_1 = \bigwedge_{v \notin \vec{\boldsymbol{u}}} \; \bigwedge_{\boldsymbol{R} \text{ is a role in } \mathcal{T}} (\boldsymbol{v} \neq \boldsymbol{c_R})$$

'all answer variables must get ABox values'

NB. if φ_1 is replaced with $\varphi_1' = \bigwedge_{v \notin \vec{\boldsymbol{u}}} \neg\text{aux}(\boldsymbol{v})$, where aux is a new relation containing all $\boldsymbol{c_R}$, then $|\boldsymbol{q}^{\dagger}| = \mathcal{O}(|\boldsymbol{q}|)$

Query Rewriting for *DL-Lite*$^{\mathcal{N}}_{horn}$ (2)

- answers to q in $\mathcal{U}_{\mathcal{K}}$ = answers to $q^{\dagger}$ in $\mathcal{I}_{\mathcal{K}}$

$\mathcal{U}_{\mathcal{K}}$ is a 'forest' model, so if t is matched to a non-ABox element then
a part of q containing t must be homomorphically embeddable into a tree

a **tree witness** $f_{R,t}$: term$(q) \to (\mathsf{N}_\mathsf{R}^-)^*$ (finite words over roles)

- $f_{R,t}(t) = \varepsilon$
- if $f_{R,t}(s) = \varepsilon$ and $R(s, s') \in q$ then $f_{R,t}(s') = R$
- if $f_{R,t}(s) = w{\cdot}S$ and $S'(s, s') \in q$ with $S' \neq S^-$ then $f_{R,t}(s') = w{\cdot}S{\cdot}S'$
- if $f_{R,t}(s) = w \cdot S$ and $S^-(s, s') \in q$ then $f_{R,t}(s') = w$

$q = \exists v\, P(v, v)$: $f_{P,v}$ does not exist

$q = \exists v_2\, (P(v_1, v_2) \land P(v_3, v_2))$: $P_{P,v_1}(v_3) = \varepsilon$

$q = \exists t_1 t_2 t_3 t_4\, (R(t_1, t_2) \land S(t_2, t_3) \land S(t_4, t_3))$:

$f_{R,t_1}(t_2) = R$, $f_{R,t_1}(t_1) = \varepsilon$, $f_{R,t_1}(t_3) = R \cdot S$, $f_{R,t_1}(t_4) = R$,
$f_{S,t_4}(t_3) = S$, $f_{S,t_4}(t_4) = \varepsilon$, $f_{S,t_4}(t_2) = \varepsilon$, $f_{S,t_4}(t_1)$ is not defined

Query Rewriting for *DL-Lite*$^{\mathcal{N}}_{horn}$ (3)

$$\varphi_2 = \bigwedge_{\substack{R(t,t')\in q \\ f_{R,t} \text{ does not exist}}} (t' \neq c_R)$$

if no tree witness exists then t cannot be mapped to a non-ABox element

$$\varphi_3 = \bigwedge_{\substack{R(t,t')\in q \\ f_{R,t} \text{ exists}}} \Big(\bigvee_{\substack{R(s,s')\in q \\ f_{R,t}(s)=\varepsilon}} (s' = c_R) \quad \rightarrow \bigwedge_{f_{R,t}(s)=\varepsilon} (s = t) \Big)$$

if both s and t are labelled with ε for role R and s' is mapped onto c_R, for $R(s, s') \in q$, then $s = t$

NB. in fact, $f_{R,t}(s) = \varepsilon$ induces an equivalence relation $\equiv^R_q$, and so, $|\varphi_3| = \mathcal{O}(|q|)$

Canonical Interpretation by FO Queries

regard the ABox as a relational instance and then

define (domain-independent) FO-queries $q_A^{\mathcal{T}}(x)$ and $q_P^{\mathcal{T}}(x,y)$ constructing $\mathcal{I}_{\mathcal{K}}$

1. for each concept C, define queries $\mathsf{exp}_C^{\mathcal{T},j}(x)$: e.g.,

(extension of concept C on step j of the SLD derivation)

$$\mathsf{exp}_A^{\mathcal{T},0}(x) = A(x)$$

$$\mathsf{exp}_C^{\mathcal{T},j+1}(x) \quad = \quad \mathsf{exp}_C^{\mathcal{T},j}(x) \quad \lor \bigvee_{C_1\sqcap\dots C_n\sqsubseteq C} \ \bigwedge_{1\leq i\leq n} \mathsf{exp}_{C_i}^{\mathcal{T},j}(x)$$

no more than $|\mathcal{T}|$ steps required

2. $q_P^{\mathcal{T}}(x,y) = P(x,y) \quad \lor \quad (\mathsf{gen}_P^{\mathcal{T}}(x) \land (y = c_P)) \quad \lor \quad (\mathsf{gen}_{P^-}^{\mathcal{T}}(y) \land (x = c_{P^-}))$

3. $q_A^{\mathcal{T}}(x) = \mathsf{exp}_A^{\mathcal{T}}(x) \land D(x)$, $\quad$ where $D(x) = \bigwedge_{c_R \in \mathsf{N}_\mathsf{I}^{\mathcal{T}}} ((x = c_R) \to \exists z\, \mathsf{gen}_R^{\mathcal{T}}(z))$

such queries can be implemented as **materialised views** (**updates!**)

Example: $\mathsf{h}(x,y) = \mathsf{h}(x,y) \lor$ h = hasTutor, t = teachesTo

$((\exists y'\, \mathsf{h}(x,y') \lor \mathsf{S}(x) \lor (x = c_t) \lor \exists y'\, \mathsf{t}(y',x)) \land \neg\exists y'\, \mathsf{h}(x,y') \land (y = c_h)) \lor$

$\exists z\, ((\exists y'\, \mathsf{t}(z,y') \lor \mathsf{P}(z) \lor (z = c_h) \lor \exists y'\, \mathsf{h}(y',z)) \land \neg\exists y'\, \mathsf{t}(z,y') \land (x = c_t) \land (y = c_h))$

Combining the two Rewriting Steps

- **polynomial** pure query rewriting for *DL-Lite*$^{\mathcal{F}}_{core}$
- and even for *DL-Lite*$^{\mathcal{N}}_{core}$ (if the aggregation function **COUNT** is available)
- otherwise $|\mathsf{exp}^{\mathcal{T},0}_{\geq k\, R}(x)| = \mathcal{O}(k^2)$,

  which is exponential in $\mathcal{T}$ if binary coding of k is used

Example:

$$q(x) = (x \neq c_h) \land (x \neq c_t) \land$$
$$\Big(\mathsf{t}(x,y) \lor ((\mathsf{P}(x) \lor \exists y'\, \mathsf{h}(y',x)) \land \neg\exists y'\, \mathsf{t}(x,y') \land (y = c_t)) \lor$$
$$\exists w\, ((\mathsf{S}(w) \lor \exists y'\, \mathsf{t}(y',w)) \land \neg\exists y'\, \mathsf{h}(w,y') \land (x = c_h) \land (y = c_t))\Big) \land$$
$$\Big(\mathsf{h}(y,z) \lor ((\mathsf{S}(y) \lor \exists z'\, \mathsf{t}(z',y)) \land \neg\exists z'\, \mathsf{h}(y,z') \land (z = c_h)) \lor$$
$$\exists w'\, ((\mathsf{P}(w') \lor \exists z'\, \mathsf{h}(z',w')) \land \neg\exists z'\, \mathsf{t}(w',z') \land (y = c_t) \land (z = c_h))\Big)$$

which is equivalent to $\quad q(x) = \mathsf{t}(x,y) \lor \mathsf{P}(x) \lor \exists y'\, \mathsf{h}(y',x)$

Other Applications of the Technique

- only exponential blowup for positive existential query answering in $\textit{DL-Lite}_{\textit{horn}}^{(\mathcal{HN})}$
- without the UNA, the technique is applicable to query answering in $\textit{DL-Lite}_{\textit{horn}}^{(\mathcal{HF})}$

  (and this is P-complete for data complexity)

- experiments show that the approach is **competitive**

  with executing the **original query** over the data

  (the formulas φ_1–φ_3 introduce additional selection conditions on top of the original query)

Open Questions

– is the exponential blowup unavoidable for role inclusions?

– is the exponential blowup unavoidable for positive existential queries?

– are there other fragments with pure polynomial rewriting?

more at `http://www.dcs.bbk.ac.uk/~roman/`