

QUERY REWRITING UNDER DL CONSTRAINTS

Héctor Pérez-Urbina, Boris Motik, and Ian Horrocks

Computing Laboratory
University of Oxford

September 25, 2008



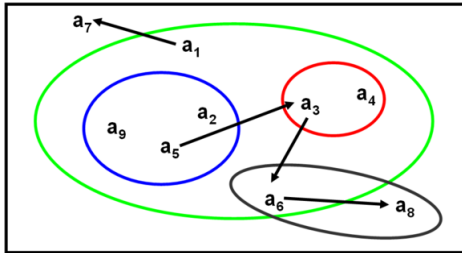


OUTLINE

- (Very brief) Introduction to Description Logics (DLs)
- Query Answering over DL Knowledge Bases
- Query Answering via Query Rewriting
- Query Rewriting for \mathcal{ELHIQ}
- Complexity Results
- Conclusion
- Future work



THE DESCRIPTION LOGIC WORLD



- A DL **Knowledge Base** (KB) is a tuple $\mathcal{K} = \langle \mathcal{T}, \mathcal{A} \rangle$
 - **TBox** $\mathcal{T} \approx$ Conceptual **schema**
 - **ABox** $\mathcal{A} \approx$ (Partial) Database **instance**



QUERYING DL KBS

- Given a n-ary query Q and a KB $\mathcal{K} = \langle \mathcal{T}, \mathcal{A} \rangle$
- $\text{ans}(Q, \mathcal{K}) = \{ \langle a_1, \dots, a_n \rangle \mid \mathcal{K} \models Q[a_1/x_1, \dots, a_n/x_n] \}$

EXAMPLE

$Q(x) \leftarrow \text{Human}(x)$

$\mathcal{T} = \{ \exists \text{hasParent.Human} \sqsubseteq \text{Human}, \text{Child} \sqsubseteq \exists \text{hasParent.Human}, \text{Father} \sqsubseteq \text{Human}, \text{hasFather} \sqsubseteq \text{hasParent} \}$

$\mathcal{A} = \{ \text{Father}(\text{JOHN}), \text{Child}(\text{MARY}), \text{hasFather}(\text{STEVE}, \text{JOHN}) \}$

- $\mathcal{T} \models \{ \text{Father} \sqsubseteq \text{Human}, \text{Child} \sqsubseteq \text{Human}, \exists \text{hasFather.Human} \sqsubseteq \text{Human} \}$
- $\text{ans}(Q, \mathcal{K}) = \{ \text{JOHN}, \text{MARY}, \text{STEVE} \}$



QUERY ANSWERING VIA QUERY REWRITING

QUERY REWRITING OVER DL TBOXES



s.t. $\text{ans}(Q, \langle \mathcal{T}, \mathcal{A} \rangle) = \text{ans}(Q', \langle \emptyset, \mathcal{A} \rangle)$, for **all** \mathcal{A}

- $Q'(x) \leftarrow \text{Human}(x) \vee \text{Child}(x) \vee \text{Father}(x) \vee \dots$



POSSIBLE FORMS OF THE REWRITING

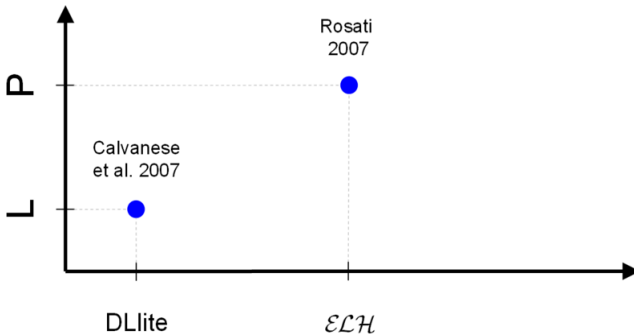
- **Fixed** query language: conjunctive queries

\mathcal{L}	Data complexity of Query Answering	Type of Rewriting
DL-Lite	LOGSPACE	Union of Conjunctive Queries (UCQs)
DL-Lite ⁺	NLOGSPACE -complete	UCQs + Linear Datalog Program
\mathcal{EL}	PTIME -complete	Datalog Program

- **No** recursion → **Linear** recursion → **Full** recursion

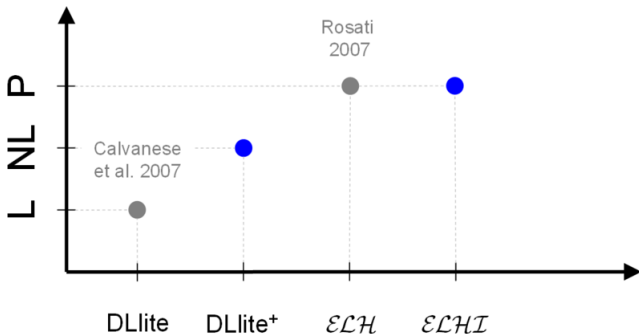


RELATED WORK



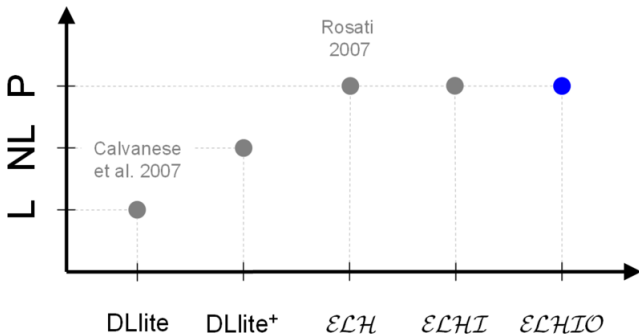


RELATED WORK CONTD.





OUR GOAL





SYNTAX

- $B \rightarrow A \mid \exists R \mid \exists R.A \mid B_1 \sqcap B_2 \mid \{a\}$
- $R \rightarrow P \mid P^-$
- TBox assertions: $B_1 \sqsubseteq B_2$ and $R_1 \sqsubseteq R_2$
- ABox assertions: $A(a)$ and $P(a, b)$

WHY BOTHER?

- $\text{Mexican} \sqsubseteq \exists \text{wasBornIn}.\{\text{MEXICO}\}$
- $\{\text{HÉCTOR MANUEL}\} \equiv \{\text{HÉCTOR}\}$

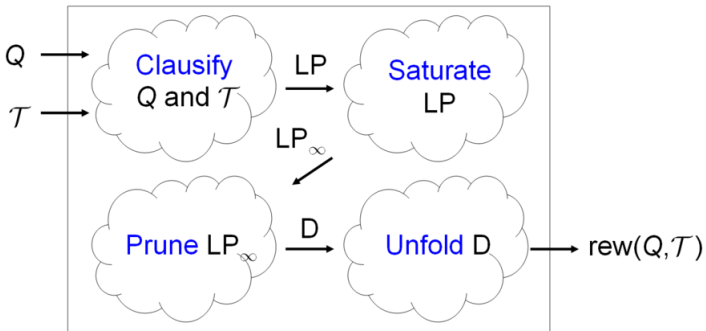


CONJUNCTIVE QUERY REWRITING OVER *ELHIO* KBS

- Worst-case **optimal** query rewriting algorithm for *ELHIO* TBoxes
- Use **known** rewriting approaches
 - Motik and Kazakov
- **Transform** Q and \mathcal{T} into a **datalog** query $\text{rew}(Q, \mathcal{T})$ using **resolution**



OUR ALGORITHM IN A NUTSHELL





CHALLENGES

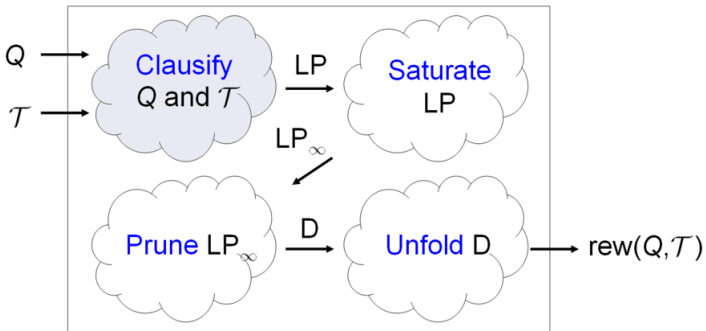
EXAMPLE

$$\mathcal{T} = \{ \text{Pope} \sqsubseteq \text{Human} \sqcap \{\text{BENEDICTXVI}\}, \\ \text{Human} \sqsubseteq \exists \text{hasFather}.\text{Human} \}$$
$$\Xi(\mathcal{T}) = \{ \text{Human}(x) \leftarrow \text{Pope}(x), \\ x \approx \text{BENEDICTXVI} \leftarrow \text{Pope}(x), \\ \text{hasFather}(x, \text{dad}(x)) \leftarrow \text{Human}(x), \\ \text{Human}(\text{dad}(x)) \leftarrow \text{Human}(x) \}$$

- Classification: Reasoning with **equality**
- Saturation: Functional terms + cycles = **trouble!**
- Pruning: Only **redundant** clauses get pruned
- Unfolding: Ensure **optimality**



CLAUSIFICATION



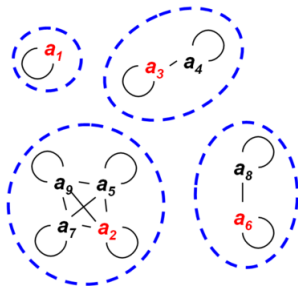


CLAUSIFICATION CONTD.

- Resolution calculi with **built-in** equality
 - Unnecessarily **complex**
 - Optimality might be **unattainable**
- **Axiomatization** of equality E_T
 - Highly **inefficient**
- **Approximate** equality: E'_T
- Obtain $\text{ans}(Q, \Xi(\mathcal{K}) \cup E_T)$ from $\text{ans}(Q, \Xi(\mathcal{K}) \cup E'_T)$



REPRESENTATIVES

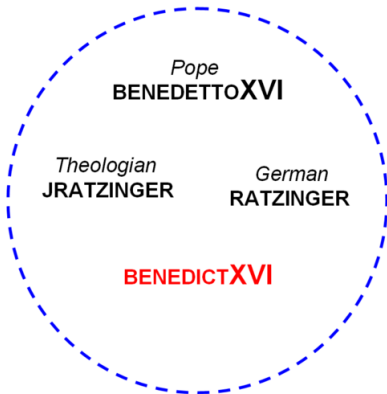


$$a_1 < a_2 < a_3 < a_4 < a_5 < a_6 < a_7 < a_8 < a_9$$

- $[a_1] = a_1$
- $[a_9] = [a_7] = [a_5] = [a_2] = a_2$
- $[a_4] = [a_3] = a_3$
- $[a_8] = [a_6] = a_6$



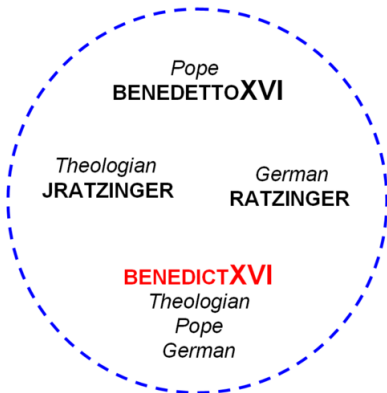
REPRESENTATIVES CONTD.



- $[BENEDICTXVI] = [JRATZINGER] = [BENEDETTOXVI] = [RATZINGER] = BENEDICTXVI$



REPRESENTATIVES CONTD.



- $[BENEDICTXVI] = [JRATZINGER] = [BENEDETTOXVI] = [RATZINGER] = BENEDICTXVI$



REPRESENTATIVES CONTD.

Careful with function symbols!

BENEDETTOXVI
JRATZINGER RATZINGER
BENEDICTXVI

dad(RATZINGER)
RATZINGERSR



REPRESENTATIVES CONTD.

Careful with function symbols!

BENEDETTOXVI
JRATZINGER RATZINGER
BENEDICTXVI

dad(RATZINGER) *dad*(JRATZINGER)
RATZINGERSR *dad*(BENEDETTOXVI)
dad(BENEDICTXVI)



APPROXIMATING EQUALITY

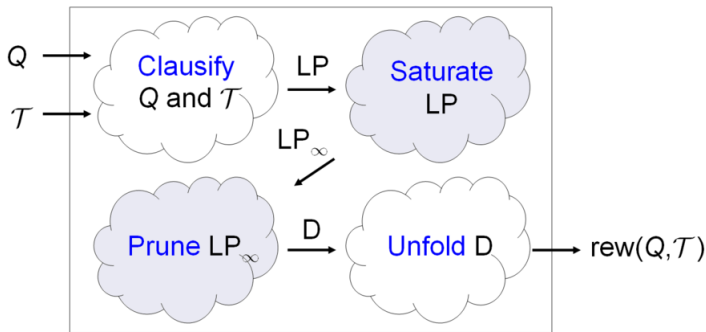
- E'_T :
 - $\mathcal{O}(o)$
 - $x \approx x \wedge \mathcal{O}(x)$
 - $x \approx y \leftarrow y \approx x \wedge \mathcal{O}(x) \wedge \mathcal{O}(y)$
 - $x \approx z \leftarrow x \approx y \wedge y \approx z \wedge \mathcal{O}(x) \wedge \mathcal{O}(y) \wedge \mathcal{O}(z)$
 - $A(y) \leftarrow A(x) \wedge x \approx y$
 - $P(x, z) \leftarrow P(x, y) \wedge y \approx z$
 - $P(z, y) \leftarrow P(x, y) \wedge x \approx z$
- $x \approx o \leftarrow A(x) \ (A \sqsubseteq \{o\})$

CORRECTNESS OF APPROXIMATION

$\langle [a_1], \dots, [a_n] \rangle \in \text{ans}(Q, \Xi(\mathcal{K}) \cup E_T)$ if and only if
 $\langle [a_1], \dots, [a_n] \rangle \in \text{ans}(Q, \Xi(\mathcal{K}) \cup E'_T)$



SATURATION AND PRUNING



- $LP = \Xi(T) \cup E'_T \cup Q$
- LP_{∞} : **saturation** of LP by \mathcal{R}^{DL}
- D: every **function-free** and **equality-free** clause in LP_{∞}



SATURATION AND PRUNING CONTD.

CHALLENGES

- Ensure **termination** of the saturation process
- Ensure **correctness** of D

- Specially tailored **selection function** for \mathcal{R}^{DL}

LEMMA

$\text{ans}(Q, \Xi(\mathcal{T}) \cup E'_{\mathcal{T}} \cup \mathcal{A}) = \text{ans}(D, \mathcal{A})$, for **every** \mathcal{A}



THE NEED FOR UNFOLDING

WHAT HAPPENS WITH DL-LITE⁺?

$$Q = \langle Q_P, \{Q_P(x) \leftarrow \text{Human}(x)\} \rangle$$

$$\mathcal{T} = \{ \exists \text{hasParent. Human} \sqsubseteq \text{Human}, \text{hasFather} \sqsubseteq \text{hasParent} \}$$

$$D = \{ \text{Human}(x) \leftarrow \text{hasParent}(y, x) \wedge \text{Human}(y), \\ \text{hasParent}(x, y) \leftarrow \text{hasFather}(x, y), Q_P(x) \leftarrow \text{Human}(x) \}$$

FROM NONLINEAR TO LINEAR DATALOG PROGRAMS

$$(1) \text{ Human}(x) \leftarrow \text{hasParent}(y, x) \wedge \text{Human}(y)$$

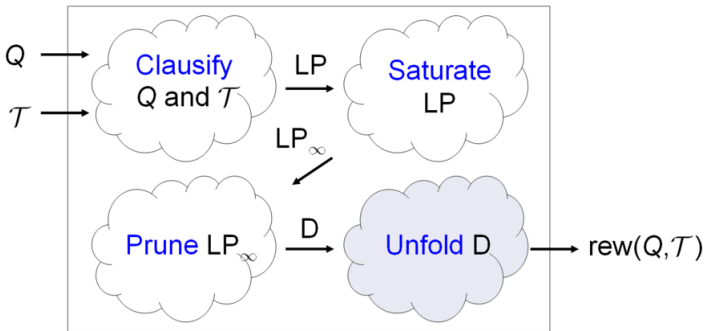
$$(2) \text{ hasParent}(x, y) \leftarrow \text{hasFather}(x, y)$$

$$(1) \text{ Human}(x) \leftarrow \text{hasParent}(y, x) \wedge \text{Human}(y)$$

$$(2) \text{ Human}(x) \leftarrow \text{hasFather}(y, x) \wedge \text{Human}(y)$$



UNFOLDING



- rew(Q, \mathcal{T}): **unfold** every clause with at most one body atom in D



UNFOLDING

EXAMPLE CONTD.

$$\text{rew}(Q, \mathcal{T}) = \{ \text{Human}(x) \leftarrow \text{hasParent}(y, x) \wedge \text{Human}(y), \\ \text{Human}(x) \leftarrow \text{hasFather}(y, x) \wedge \text{Human}(y), \\ Q_P(x) \leftarrow \text{Human}(x) \}$$

THEOREM

$$\text{ans}(Q, \Xi(\mathcal{T}) \cup E'_T \cup \mathcal{A}) = \text{ans}(\text{rew}(Q, \mathcal{T}), \mathcal{A}), \text{ for every } \mathcal{A}$$



EXAMPLE

TBOX AND QUERY

$\mathcal{T} = \{ \text{Pope} \sqsubseteq \text{Human} \sqcap \{ \text{BENEDICTXVI} \},$
 $\text{Human} \sqsubseteq \exists \text{hasParent.Human},$
 $\text{hasFather} \sqsubseteq \text{hasParent} \}$

$Q = Q(x) \leftarrow \text{Human}(x)$

CLAUSIFICATION: LP

$\text{Human}(x)$	$\leftarrow \text{Pope}(x),$
$x \approx \text{BENEDICTXVI}$	$\leftarrow \text{Pope}(x),$
$\text{hasParent}(x, \text{parent}(x))$	$\leftarrow \text{Human}(x),$
$\text{Human}(\text{parent}(x))$	$\leftarrow \text{Human}(x),$
$\text{hasParent}(x, y)$	$\leftarrow \text{hasFather}(x, y),$
$Q(x)$	$\leftarrow \text{Human}(x),$

plus E'_T



EXAMPLE CONTD.

SATURATION AND PRUNING: D

$Human(x)$ $\leftarrow Pope(x),$
 $Pope(BENEDICTXVI)$ $\leftarrow Pope(x),$
 $Q(x)$ $\leftarrow Human(x)$

UNFOLDING: $rew(Q, \mathcal{T})$

$Q(x)$ $\leftarrow Human(x)$
 $Q(x)$ $\leftarrow Pope(x)$
 $Q(BENEDICTXVI)$ $\leftarrow Pope(x)$



PROPERTIES OF THE REWRITING

LEMMA

If \mathcal{T} is in \mathcal{ELHIO} , then $\text{rew}(Q, \mathcal{T})$ is a **datalog program**

LEMMA

If \mathcal{T} is in DL-Lite^+ , then $\text{rew}(Q, \mathcal{T})$ consists of a **union of conjunctive queries** and a **linear datalog query**

LEMMA

If \mathcal{T} is in DL-Lite , then $\text{rew}(Q, \mathcal{T})$ is a **union of conjunctive queries**



COMPLEXITY ANALYSIS

THEOREM

For a **conjunctive query** Q and a \mathcal{ELHIO} KB $\mathcal{K} = \langle \mathcal{T}, \mathcal{A} \rangle$, deciding whether $\vec{a} \in \text{ans}(Q, \mathcal{K})$ is **PTIME-complete** w.r.t. data complexity

COROLLARY

Given a **conjunctive query** Q and a KB $\mathcal{K} = \langle \mathcal{T}, \mathcal{A} \rangle$, if \mathcal{T} is in \mathcal{ELHI} , DL-Lite^+ , or DL-Lite we can compute the answers to $\text{rew}(Q, \mathcal{T})$ over \mathcal{A} in **PTIME**, **NLOGSPACE**, or **LOGSPACE** respectively w.r.t. data complexity

- $\text{rew}(Q, \mathcal{T})$ is **optimal!**



CONCLUSION

NEW RESULT

QA over \mathcal{ELHIO} KBs is **PTIME-complete** w.r.t. data complexity

TAKE HOME MESSAGE

Conjunctive query **rewriting** algorithm for \mathcal{ELHIO} KBs

- Worst-case **optimal** for sublanguages of \mathcal{ELHIO} for which QA is **PTIME-complete**, **NLOGSPACE-complete**, or in **LOGSPACE** w.r.t. data complexity
- **Generalization** and **extension** of existing approaches
- Use of (deductive) **DB technology** for query evaluation



CURRENT AND FUTURE WORK

CURRENT WORK

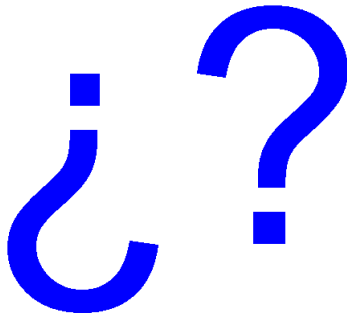
- Technical report
- ReQuEM (Rewriting Queries In Expressive Models)

FUTURE WORK

- Optimization
- Information Integration



THANKS!





AXIOMATIZATION OF EQUALITY E_T

- $x \approx x$
- $x \approx y \leftarrow y \approx x$
- $x \approx z \leftarrow x \approx y \wedge y \approx z$
- $f(x) \approx f(y) \leftarrow x \approx y$
- $A(y) \leftarrow A(x) \wedge x \approx y$
- $P(x, z) \leftarrow P(x, y) \wedge y \approx z$
- $P(z, y) \leftarrow P(x, y) \wedge x \approx z$