Data Complexity of Query Answering in Description Logics
(Extended Abstract)

D. Calvanese1, G. De Giacomo2, D. Lembo2, M.Lenzerini2, R. Rosati2

1 Faculty of Computer Science
Free University of Bozen-Bolzano
calvanese@inf.unibz.it

2 Dip. di Ing. Informatica, Automatica e Gestionale
Sapienza Università di Roma
lastname@dis.uniroma1.it

Abstract

We study the data complexity of answering conjunctive queries over Description Logic knowledge bases constituted by a TBox and an ABox. In particular, we are interested in characterizing the FO-rewritability and the polynomial tractability boundaries of conjunctive query answering, depending on the expressive power of the DL used to express the knowledge base. What emerges from our complexity analysis is that the Description Logics of the DL-Lite family are essentially the maximal logics allowing for conjunctive query answering through standard database technology.

1 Introduction

The idea of using ontologies as a conceptual view over data repositories is becoming more and more popular. For example, in Enterprise Application Integration [Lee et al., 2003], Data Integration [Lenzerini, 2002], and the Semantic Web [Heflin and Hendler, 2001], the intensional level of the application domain can be profitably represented by an ontology, so that clients can rely on a shared conceptualization when accessing the services provided by the system. In these contexts, the set of instances of the concepts in the ontology is to be managed in the data layer of the system architecture (e.g., in the lowest of the three tiers of the Enterprise Software Architecture), and, since instances correspond to the data items of the underlying information system, such a layer constitutes a repository that is very large (much larger than the intensional level of the ontology), to be stored in secondary storage (see, e.g., [Borgida et al., 1989]).

When clients access the application ontology, it is very likely that one of the main services they need is the one of answering complex queries over the extensional level of the ontology, which means computing the query answers that are logically implied by the whole ontology. Here, by ‘complex’ we mean that it does not suffice to ask for the instances of concepts, but we need at least to express conjunctive conditions on the extensional level [Calvanese et al., 1998; Horrocks and Tessaris, 2000; Fikes et al., 2005; Calvanese et al., 2007b; 2008; Lutz, 2008; Glimm et al., 2008]. Given the size of the instance repository, when measuring the computational complexity of query answering (and reasoning in general) the most important parameter is the size of the data, i.e., we are interested in the so-called data complexity of query answering [Vardi, 1982].

We consider here conjunctive queries (CQs) specified over ontologies expressed in Description Logics (DLs), and study the data complexity of query answering. Since an ontology in DLs is essentially a knowledge base (KB) constituted by a TBox and an ABox, the problem we address is the one of computing the answers to a CQ that are logical consequences of the TBox and the ABox, where complexity is measured with respect to the size of the ABox only. Note that we borrow the notion of data complexity from the database literature [Vardi, 1982], on the premise that an ABox can be naturally viewed as a relational database. Recently, data complexity has attracted the interest of the DL community, first for reasoning over TBox and ABox (i.e., instance checking, which is the simplest form of query answering) [Donini et al., 1994; Hustadt et al., 2005], and then also for answering full conjunctive queries [Ortiz et al., 2008; Eiter et al., 2009]. This gave rise to the study of DLs for which query answering can be done efficiently in data complexity [Calvanese et al., 2005; 2007a; Krishnadhi and Lutz, 2007; Artale et al., 2009], which is a key aspect of the present paper.

Specifically, we are interested in characterizing the FO-rewritability and the polynomial tractability boundaries of conjunctive query answering, depending on the expressive power of the DL used to specify the KB. We say that query answering is FO-rewritable in a DL L, if for every conjunctive query q over an L TBox T, one can effectively compute a first-order (FO) query qL over L such that the answers to qL over T are the same as the answers to qL over the database corresponding to the ABox A. Since first-order queries can be expressed in SQL, the importance of FO-rewritability is that, when query answering enjoys this property, we can take advantage of Relational Data Base Management System (RDBMS) techniques for both representing data, i.e., ABox assertions, and answering queries via reformulation into SQL. Notably, in this case, the data complexity of conjunctive query answering over ontologies is the one of evaluating FO queries over relational databases, i.e., AC0 [Abiteboul et al., 1995], a
complexity class strictly contained in LOGSPACE [Reingold, 2008]. In the class of DLs for which query answering is FO-rewritable, we essentially find the DLs of the DL-Lite family [Calvanese et al., 2007a]. Notably, the two simplest DLs of this family (namely, DL-LiteR and DL-LiteF) are rich enough to express basic ontology languages, e.g., extensions of (the DL subset of) RDFS, or fragments of OWL 2; conceptual data models, e.g., Entity-Relationship; and object-oriented formalisms, e.g., basic UML class diagrams. In fact, in the present paper we consider a new DL of the DL-Lite family, called DLR-LiteA,F, which generalizes both DL-LiteR and DL-LiteF by allowing for the use of n-ary relations between (instances of) concepts, the specification of keys on relations, combined together (in a controlled way) with inclusions between (projections on) relations, and the use of conjunctions in the left-hand side of the inclusion assertions constituting the knowledge base TBox.

We are also interested in knowing for which DLs we go beyond FO-rewritability. For this purpose, we single out those DLs for which query answering becomes NLOGSPACE-hard and PTIME-hard, respectively, thus not allowing for FO-rewritability. In spite of the fact that for such languages query answering is polynomially tractable (in NLOGSPACE and PTIME, respectively), these hardness results tell us that for query answering we cannot take advantage of state-of-the-art database query optimization strategies, and this might hamper practical feasibility for very large ABoxes.

Finally, we address the problem of going even beyond PTIME, establishing coNP-hardness of conjunctive query answering with respect to data complexity for surprisingly simple DLs. In particular, we show that we get intractability as soon as the DL is able to express simple forms of union.

What emerges from our complexity analysis is that the DLs of the DL-Lite family enjoy FO-rewritability of conjunctive query answering and cannot be extended with any construct typical of DLs [Baader et al., 2003] without losing this property. In this sense, the DLs of the DL-Lite family studied here are the maximal logics that allow for answering conjunctive queries through standard database technology.

This paper is an extend abstract of [Calvanese et al., 2013]. For a complete treatment, formal proofs of all results, and a thorough discussion of related work we refer to the full paper.

2 Preliminaries

Description Logics (DLs) [Baader et al., 2003] are logics that represent the domain of interest in terms of objects, i.e., individuals, concepts, which are abstractions for sets of objects, and relations among concepts. Relations are typically binary in DLs (they are called roles), but in this paper we also consider n-ary relations, in the spirit of the DL DLR [Calvanese et al., 1998; 2008].

A DL knowledge base (KB) \( K = < T, A > \) is a pair formed by a set \( T \), called TBox, of intensional assertions, i.e., axioms specifying general properties of concepts, and a set \( A \), called ABox, of extensional assertions, i.e., axioms about individual objects. Informally, in an ABox a concept assertion specifies that an object is an instance of an atomic concept. Analogously, the other types of ABox assertions specify instances of atomic roles and relations.

Different DLs allow for both different concept and role expressions, and different TBox intensional assertions. In other words, defining a specific DL means providing a specification of both the language for building complex expressions, and the language for specifying intensional assertions.

In this paper we consider the following constructs:

\[
\begin{align*}
C & \rightarrow A \mid \neg C \mid C \cap \cdots \cap C \mid C \cup \cdots \cup C \\
Q & \rightarrow P \mid P^\bot \mid \neg P \mid \neg P^\bot \\
V & \rightarrow R \mid R^\ell[i_1, \ldots, i_h] \mid \neg R \mid \neg R^\ell[i_1, \ldots, i_h],
\end{align*}
\]

where \( A \) denotes an atomic concept, \( P \) an atomic role, and \( P^\bot \) its inverse, \( \neg \), \( \cap \), \( \cup \) are the usual Boolean constructs, and \( R \) is an atomic n-ary relation. \( \exists Q.C \) is the qualified existential role quantification denoting the objects that have a Q-role successor that belongs to \( C. \exists Q.C \) is its unqualified variant denoting simply the objects that have a Q-role successor. \( \forall Q.C \) is the universal quantification, denoting the objects whose Q role successors are all in \( C. \exists i:R \) denotes the objects that participate as i-th component of the n-ary relation \( R \). Finally \( R^\ell[i_1, \ldots, i_h] \) is the relation corresponding to the projection of \( R \) on its \( i_1, \ldots, i_h \) components.

Table 1 illustrates the various TBox and ABox assertions used in this paper. We distinguish between the constructs that we allow in the left-hand side (\( C, Ql, VI \)) and in the right-hand side (\( Cr, Qr, Vr \)) of inclusion assertions. As TBox assertions we have concept, role, and relation inclusions, expressing containment between sets, pairs, and tuples of objects, respectively. We then have role functionality assertions expressing that a (direct or inverse) role is a functional binary relation, and key assertions expressing that there are no two tuples of objects in a relation \( V \) sharing the same key components \( j_1, \ldots, j_l \). As ABox assertions, we have ground facts involving atomic concepts, roles, and relations.

The semantics of a DL KB is given in terms of first-order interpretations. We say that one such interpretation \( I \) is a model of a KB \( K = < T, A > \) if \( I \) satisfies all assertions in \( T \) and in \( A. \) \( K \) is satisfiable if it has at least a model. Given a sentence \( \phi, K \models \phi \) denotes the implication of \( \phi \) by \( K \), i.e., the fact that \( \phi \) holds in all models of \( K \).

Example 1. Let us assume that our signature includes the atomic concepts Supplier, Customer, and Product, the ternary relation supply, and the binary relation clientOf. We define the following TBox \( T_e \):
\[\exists 1: \text{supplier} \sqsubseteq \text{Supplier} \quad (1)\]
\[\exists 2: \text{supplier} \sqsubseteq \text{Customer} \quad (2)\]
\[\exists 3: \text{supply} \sqsubseteq \text{Product} \quad (3)\]
\[\text{Product} \sqsubseteq \neg \text{Supplier} \quad (4)\]
\[\text{Product} \sqsubseteq \neg \text{Customer} \quad (5)\]
\[\text{(key 2, 3: supply)} \quad (6)\]
\[\text{Supplier} \sqcap \text{Customer} \sqsubseteq \exists 1: \text{supply} \quad (7)\]
\[\text{Supplier} \sqcap \text{Customer} \sqsubseteq \exists 2: \text{supply} \quad (8)\]
\[\text{supply} [1, 2] \sqsubseteq \text{clientOf} [2, 1] \quad (9)\]

In the above TBox, inclusions (1)–(3) specify the domain respectively of the first, second, and third component of the relation supply, with the intended meaning that suppliers provide customers with products. Assertions (4) and (5) impose that the set of products is disjoint from the set of customers and the set of suppliers, respectively. Assertion (6) imposes that positions 2 and 3 in supply constitute a key of supply, with the intended meaning that a customer for a certain product has only one supplier. Assertions (7) and (8) specify that those individuals that are both suppliers and customers must participate in both the first and the second component of the relation supply. Finally, assertion (9) says that each individual that is a supplier of a customer (for a certain product), has such a customer as a client. As an ABox, consider:
\[A_e = \{ \text{Customer}(\text{SmithInc}), \text{Supplier}(\text{SmithInc}), \text{clientOf}(\text{SmithInc}, \text{SmartCompany}) \}\]

We consider conjunctive queries (CQs) over satisfiable DL KBs. Given one such query \(q\) over a KB \(K = (T, A)\), answering \(q\) over \(K\) amounts to computing its certain answers over \(K\), i.e., all the tuples \(\vec{t}\) of constants in \(K\) such that \(K \models q(\vec{t})\), where \(q(\vec{t})\) denotes the query obtained from \(q\) by replacing its free variables with the constants in \(\vec{t}\).

We say that (conjunctive) query answering in a DL language \(\mathcal{L}\) is FO-rewritable, if for every TBox \(T\) expressed in \(\mathcal{L}\) and every (conjunctive) query \(q\) over \(T\), one can effectively compute a FO query \(q_0\) over \(T\) such that the following holds: for every ABox \(A\) for which \((T, A)\) is satisfiable, and every tuple \(t\) of constants occurring in \(A\), \((T, A) \models q(t)\) if and only if \((0, A) \models q_0(t)\). The query \(q_0\) is called the FO-rewriting of \(q\) w.r.t. \(T\). In this paper, we study data complexity of query answering, i.e., the complexity of the decision problem associated to query answering measured w.r.t. the size of the ABox only, i.e., by considering both the input TBox and the input query as fixed. Observe that every FO query can be evaluated in \(AC^0\) with respect to data complexity (see e.g., [Abiteboul et al., 1995]). It follows that, if query answering in \(\mathcal{L}\) is FO-rewritable, then query answering in \(\mathcal{L}\) is in \(AC^0\) w.r.t. data complexity. Vice-versa, if query answering (or KB satisfiability) is \(C\)-hard w.r.t. data complexity for some complexity class \(C\) that strictly contains \(AC^0\) (e.g., LOGSPACE, NLOGSPACE, PTIME, coNP, etc.), then it is not FO-rewritable.

### 3 DLR-Lite\(_{A,\neg}\) and FO-rewritability

The logic DLR-Lite\(_{A,\neg}\) is one the most expressive members of the DL-Lite family. It extends the basic DL-Lite logics DLR-Lite\(_F\) and DLR-Lite\(_X\) [Calvanese et al., 2007a] with \(n\)-ary relations, inclusions between projections of \(n\)-ary relations, conjunction in the left-hand side of concept inclusions, and key dependencies on (projections of) \(n\)-ary relations (in a controlled way). The syntax of DLR-Lite\(_{A,\neg}\) is given in Table 2. The KB given in Example 1 is a DLR-Lite\(_{A,\neg}\) KB.

In [Calvanese et al., 2013], we show that query answering in DLR-Lite\(_{A,\neg}\) is FO-rewritable, and therefore that it is in \(AC^0\) in data complexity. To this aim we provide an algorithm, called PerfectRef, that, taken a CQ \(q\) and a DLR-Lite\(_{A,\neg}\) TBox \(T\) as input, returns a FO query \(q_0\), in fact a union of conjunctive queries, that we prove to be a FO-rewriting of \(q\) w.r.t. \(T\). In a nutshell, the algorithm compiles in \(q_0\) both the query \(q\) and the assertions of \(T\) that are relevant to compute the answers to \(q\). Notably, to compute \(q_0\), PerfectRef needs to consider only the positive inclusions explicitly asserted in \(T\), i.e., inclusions that do not use negation. Such inclusions are used as rewriting rules, applied from right to left, for the query \(q\). Besides rewriting steps, PerfectRef performs also some unifications between query atoms, which may make further rewriting steps applicable. Both such steps are executed iteratively until a fixpoint is reached. We refer to [Calvanese et al., 2013] for further details. We notice that the algorithm we described is structurally similar to an analogous algorithm presented in [Calvanese et al., 2007a] for computing the FO-rewriting of a CQ over either a DLR-Lite\(_F\) or a DLR-Lite\(_X\) TBox. The PerfectRef algorithm we discuss here is however complicated by the fact that while rewriting the query, it has to properly manage the presence of \(n\)-ary relations, both in the query and in the inclusions, and the presence of conjunctions in concept inclusions. In particular, such conjunctions may cause some CQs produced by the algorithm as part of the rewriting to have more atoms than the original input query (notice that for DL-Lite logics without conjunctions in the left-hand side of inclusions, all rewritten CQs have at most as many atoms as the input query).

**Example 2.** Consider the query \(q(x) \leftarrow \text{supply}(x, y, z), \text{Product}(z)\), posed over the TBox \(T_0\) of Example 1. The algorithm applies inclusion (3) and generates the query \(q(x) \leftarrow \text{supply}(x, y, z), \text{supply}(w_1, w_2, z)\). Such a query cannot be further rewritten through positive inclusions. Indeed, the only inclusions having \(\text{supply}\) in their right-hand side, i.e., inclusions (7) and (8), do not propagate all join and free variables occurring in the query atoms. By unifying the two
supply atoms in the query, with unifier \{w_1/x, w_2/y\}, the algorithm then produces the query \(q(x) \leftarrow \text{supply}(x, y, z)\).

Such a query can be rewritten using inclusion (7), and the query \(q(x) \leftarrow \text{Supplier}(x), \text{Customer}(x)\) is thus added to the rewriting. We notice that the unification step is necessary to generate this query. The evaluation of the last query over the ABox \(A_e\) produces the set \{SmithInc\}. Such a set constitutes in fact the set of certain answers of the input query over the KB \(\langle T_e, A_e \rangle\).

4 Going beyond FO-rewritability

Next, we show that, as soon as we consider further, minimal extensions of \(\text{DLR-Lite}_{A, r}\), we cross the boundary of \(\text{AC}^0\) data complexity. Going beyond \(\text{AC}^0\) data complexity means actually that we lose the property of FO-rewritability and therefore query answering requires more powerful engines than those available in standard relational database technology. An immediate consequence of this fact is that we cannot take advantage anymore of data management tools and query optimization techniques of current DBMSs.

We point out that the extensions of \(\text{DLR-Lite}_{A, r}\) that we consider make query answering harder even also if we restrict relations to be binary, i.e., if we have only roles. Therefore, in the following, we consider DLs with roles rather than n-ary relations. Moreover, such extensions have a negative impact on complexity even if we consider them alone or added to very basic DLs. Specifically, we study the complexity of query answering for DLs in which the ABox of a KB is as to very basic DLs. Specifically, we study the complexity of query answering has become a central theme in properties, in particular FO-rewritability and thus \(\text{AC}^0\) complexity of query answering. We have then identified rather new types of DLs for which query answering is no longer reducible to evaluation of a first-order logic formula (and hence an instance checking). (*) Relations in key assertions are not specialized.

In our investigation we identify various DL languages for which conjunctive query answering is \(\text{NLOGSPACE}\)-hard, \(\text{PTIME}\)-hard, or \(\text{coNP}\)-hard.

As for \(\text{NLOGSPACE}\)-hard DLs, we notice that to reach such complexity it suffices to allow for the use of existen- tial role quantification qualified on atomic concepts in the left-hand side of positive concept inclusions (cf. row 2 in Table 3). Indeed, this kind of TBox assertion requires per se the use of recursion to answer a CQ. To get an intuition, consider for example the assertion \(\exists \text{FatherOf} \sqsubseteq \text{Human} \sqsubseteq \text{Human}\), which states that each individual that is father of a human is also a human. Then, in order to answer a query for all individuals that are humans, we need to retrieve those that are asserted to be human, those that are fathers of asserted humans, those that are fathers of fathers of asserted humans, and so on. This computation requires linear recursion. In [Calvanese et al., 2013] we indeed prove \(\text{NLOGSPACE}\)-hardness for this case by a \(\text{LOGSPACE}\) reduction from reachability in directed graphs. For the logic in row 3 in Table 3, \(\text{NLOGSPACE}\)-hardness is proved by showing that every TBox in this logic is logically equivalent to a TBox expressed in the logic of row 2. As for the logic in row 4, we exhibit a different reduction from reachability in directed graphs. Notice that for all such DLs we need (at least) linear Datalog to answer CQs.

To prove \(\text{PTIME}\)-hardness results, one can use \(\text{LOGSPACE}\) reductions from path system accessibility, see [Calvanese et al., 2013]. Note that \(\text{PTIME}\)-hardness implies that we need at least the power of full Datalog to answer CQs in these cases.

The intuition behind the \(\text{coNP}\)-hardness results given in Table 3 is that in all three cases we consider it is possible to require a reasoning by case analysis, caused by set covering assertions. Indeed, whereas in the first \(\text{coNP}\)-hard logic (cf. row 9) we can explicitly assert set covering through the use of disjunction in \(C_r\), for the two other cases covering can be asserted on the entire domain. More precisely, in the second \(\text{coNP}\)-hard logic (cf. row 10), the domain can be covered by \(A_1\) and \(A_2\) through an assertion of the form \(\neg A_1 \sqsubseteq A_2\), whereas in the last case (cf. row 11), the domain can be implicitly covered by existentials in the CQ and an assertion of the form \(\forall P \cdot A_1 \sqsubseteq A_2\), see [Calvanese et al., 2013].

Finally, note that all our \(\text{NLOGSPACE}\)-hardness and \(\text{PTIME}\)-hardness results hold already for instance checking, a simpler form of query answering where the query is a ground atom. All our complexity bounds are tight (‘c’(omplete) in Table 3) [Calvanese et al., 2013], except for the DL in row 4, for which \(\text{NLOGSPACE}\)-membership is open.

5 Conclusions

In this paper, we have presented \(\text{DLR-Lite}_{A, r}\), an interesting \(n\)-ary DL belonging to the \(\text{DL-Lite}\) family, which can be considered as the relation oriented version of the well known logic \(\text{DL-Lite}_{A}\), and which enjoys nice computational properties, in particular FO-rewritability and thus \(\text{AC}^0\) data complexity of query answering. We have then identified rather simple DLs for which query answering is no longer reducible to evaluation of a first-order logic formula (and hence an SQL query) over the data. The results provided in this paper are summarized in Table 3. We observe that data complexity of query answering has become a central theme in the current DL research, and specific attention is now given to FO-rewritability in designing languages for ontology applications that need to deal with large quantities of data [Artale et al., 2009; Ortiz et al., 2011; Lutz and Wolter, 2012; Stefanoni et al., 2014; Bienvenu et al., 2014].
References


