Information and Computation 237 (2014) 12–55

Contents lists available at ScienceDirect

Information and Computation

www.elsevier.com/locate/yinco

Answering regular path queries in expressive Description Logics via alternating tree-automata ☆

Diego Calvanese [a], Thomas Eiter [b], Magdalena Ortiz [b,*]

[a] KRDB Research Centre, Free University of Bozen-Bolzano, Via della Mostra 4, I-39100 Bolzano, Italy
[b] Institute of Information Systems, Vienna University of Technology, Favoritenstraße 9-11, A-1040 Vienna, Austria

ARTICLE INFO

Article history:
Received 23 December 2009
Received in revised form 3 April 2013
Available online 18 April 2014

Keywords:
Expressive Description Logics
Query answering
Computational complexity
Automata on infinite trees

ABSTRACT

Expressive Description Logics (DLs) have been advocated as formalisms for modeling the domain of interest in various application areas, including the Semantic Web, data and information integration, peer-to-peer data management, and ontology-based data access. An important requirement there is the ability to answer complex queries beyond instance retrieval, taking into account constraints expressed in a knowledge base. We consider this task for positive 2-way regular path queries (P2RPQs) over knowledge bases in the expressive DL $\mathcal{ZIQ}$. P2RPQs are more general than conjunctive queries, union of conjunctive queries, and regular path queries from the literature. They allow regular expressions over roles and data joins that require inverse paths. The DL $\mathcal{ZIQ}$ extends the core DL $\mathcal{ALC}$ with qualified number restrictions, inverse roles, safe Boolean role expressions, regular expressions over roles, and concepts of the form $\exists S.\mathsf{Self}$ in the style of the DL $\mathcal{SRIQ}$. Using techniques based on two-way tree-automata, we first provide as a stepping stone an elegant characterization of TBox and ABox satisfiability testing which gives us a tight ExpTime bound for this problem (under unary number encoding). We then establish a double exponential upper bound for answering P2RPQs over $\mathcal{ZIQ}$ knowledge bases; this bound is tight. Our result significantly pushes the frontier of 2ExpTime decidability of query answering in expressive DLs, both with respect to the query language and the considered DL. Furthermore, by reducing the well known DL $\mathcal{SRIQ}$ to $\mathcal{ZIQ}$ (with an exponential blow-up in the size of the knowledge base), we also provide a tight 2ExpTime upper bound for knowledge base satisfiability in $\mathcal{SRIQ}$ and establish the decidability of query answering for this significant fragment of the new OWL 2 standard.

1. Introduction

Description Logics (DLs) [2] is a well-established branch of logics for knowledge representation and reasoning, and today the premier logic-based formalisms for modeling concepts (i.e., classes of objects) and roles (i.e., binary relationships between classes). It has gained increasing attention in different areas including the Semantic Web, data and information integration, peer-to-peer data management, and ontology-based data access. In particular, many of the standard Web ontologies from the OWL family are based on DLs: the new OWL 2 standard [3] is based on a DL known as $\mathcal{SROIQ}$ [4], whose fragment $\mathcal{SRIQ}$ [5] extends the DL $\mathcal{SHIQ}$ underlying OWL-Lite [6].

☆ Some results of this paper have appeared, in preliminary form, in a conference paper in the *Proc. of AAAI 2007* [1].

* Corresponding author.
E-mail addresses: calvanese@inf.unibz.it (D. Calvanese), eiter@kr.tuwien.ac.at (T. Eiter), ortiz@kr.tuwien.ac.at (M. Ortiz).

http://dx.doi.org/10.1016/j.ic.2014.04.002
0890-5401/

In DLs, reasoning tasks like classification and instance checking, which deal with taxonomic issues, had been traditionally studied. However, the widening range of applications in which DLs are used has motivated an increasing interest in query languages whose expressive power goes beyond that of DL concept and role expressions. The aim of such languages is to allow one to join pieces of information in finding the query answer, thus overcoming one of the most significant drawbacks of DLs as languages for data management. Since the initial work of Calvanese et al. [7], many further works have addressed the problem of evaluating complex queries over DL knowledge bases. Special attention has been devoted to (*unions of*) *conjunctive queries* (CQs and UCQs) [8], which are the formal counterpart of the most widely used fragments of SQL (or relational algebra) queries, namely (unions of) select-project-join queries. (U)CQs over DL knowledge bases have been studied for many DLs, ranging from weak ones that allow for efficient algorithms, like those of the $\mathcal{EL}$ [9–11] and DL-Lite families [12], to the very expressive ones of the $\mathcal{ALCH}$ and $\mathcal{SH}$ families, cf. [13–16].

Another important language for querying knowledge bases is that of *regular path queries* (RPQs) [17–19], which allow one to ask for pairs of objects that are connected by a path conforming to a regular expression. Due to their capability of expressing complex navigations in graphs, RPQs are the fundamental mechanism for querying semi-structured and graph-structured data. Indeed, as a query language, RPQs go beyond first-order logic, since they allow one to express a *controlled* form of recursion. This turns out to be essential for querying graph-like structures such as those encountered in several domains that are gaining increasing importance, notably social networks [20] and linked open data [21]. Notice that, in a setting of incomplete information as the one encoded by means of a knowledge base, the use of unrestricted recursion would quickly lead to undecidability, not only of intensional inference tasks such as query containment or equivalence [22], but also of query answering [23,24]. Instead, the restricted form of recursion provided by RPQs and their extensions considered here, provides a good trade-off between the ability to flexibly traverse data whose precise structure is not defined a priori (e.g., in terms of a relational schema), and the decidability of query answering also in the presence of complex domain knowledge encoded in DLs. The complex paths allowed in the query allow one to find in the data complex relations between items, without being constrained by the relations explicitly stated in the data or pre-defined in the ontology, and without having to modify the ontology solely for query answering. Moreover, when also *inverse roles* are allowed to occur in the regular expression, the complex relations expressed by the resulting *two-way RPQs* (2RPQs) are not restricted to the direction initially chosen by the designer to represent relations between data items. 2RPQs are for example present in the property paths in SPARQL 1.1 [25], the new standard RDF query language, and in the XML query language XPath [26]. We consider here the yet more expressive class of *positive (existential) two-way regular path queries* (in short, P2RPQs), which are inductively built using conjunction and disjunction from atoms that are regular expressions over direct and inverse roles and allow for testing the objects encountered during navigation for membership in concepts. P2RPQs, which subsume CQs and unions of CQs, are also a natural generalization of several extensions of RPQs that have been studied by different authors, e.g., [27,28,19,29–32]. They are, to our knowledge, the most expressive query language considered so far over DL knowledge bases [33,1].

In this paper, we describe a technique, first presented in [1], for deciding the entailment problem for P2RPQs expressed over $\mathcal{ZIQ}$ knowledge bases. In query entailment, we are given a knowledge base and a Boolean query, i.e., a query that in a given interpretation evaluates either to true or to false, expressed over that knowledge base, and we are asked to determine whether the query evaluates to true in every model of the knowledge base. The DL $\mathcal{ZIQ}$, also known as $\mathcal{ALCQIb}^{\mathsf{Self}}_{reg}$, extends the well known DL $\mathcal{ALCQIb}$ (to which reasoning in $\mathcal{SHIQ}$ can be reduced [34]) with regular role expressions [35], Boolean role inclusion axioms, and concepts of the form $\exists S.\mathsf{Self}$ [5]. By means of a translation that reduces the query entailment problem over $\mathcal{SRIQ}$ KBs to $\mathcal{ZIQ}$ KBs, we also obtain an algorithm for entailment of P2RPQs over $\mathcal{SRIQ}$ knowledge bases. This is the first algorithm for query entailment (and hence for query answering) that allows both for regular expressions and for conjunctions of atoms in the query, while considering, on the DL side, a logic that extends $\mathcal{ALC}$ with inverses and counting and, notably, also supports the kind of complex role inclusions that have been advocated in the new OWL standards [3].

Previously, algorithms for query answering over expressive DLs had used a variety of techniques, ranging from query rewriting [7,13,36], over modified tableaux techniques [16], to resolution [37]. We obtain our results by exploiting techniques based on *automata on infinite trees* [38], which have been developed initially in the context of modal logics and program logics [39–43]. While the application of automata techniques in DLs is not novel, cf. [35,44,45], previous work was concerned with deciding satisfiability of a knowledge base consisting of a taxonomy part (TBox) only. Here we address the much more involved task of query answering over a knowledge base, which also has a data part (an ABox). Specifically, we extend previous automata-based algorithms for TBox satisfiability [35,44] and incorporate the ABox part. Then, to decide query entailment over DL knowledge bases, we build on the ideas of Calvanese et al. [30], which had been developed in the context of automata on finite words, and extend them to automata over infinite trees. For deciding query entailment, we implement automata operations that rely on transformations between different kinds of automata, which, from a technical point of view, are more challenging in our case than in the case of finite words. The technique we present here has been recently extended to some DLs that support nominals [33].

In this paper, we make the following contributions (all complexity results hold under unary number encoding):

- As a stepping stone to our main results, we first present an automata-based algorithm for checking the satisfiability of a $\mathcal{ZIQ}$ knowledge base that comprises both a TBox and an ABox, and that runs in ExpTime, which is worst-case optimal.