A Plausibility Description Logics for Reasoning with Information Sources Having Different Formats and Structures

Luigi Palopoli, Giorgio Terracina, and Domenico Ursino

Dipartimento di Informatica, Matematica, Elettronica e Trasporti
Università degli Studi "Mediterranea" di Reggio Calabria
Via Graziella, Località Feo di Vito, 89060 Reggio Calabria, Italy
{palopoli,terracina,ursino}@ing.unirc.it

Abstract. The aim of this paper is to illustrate how a probabilistic Description Logics, called DL_P, can be exploited for reasoning about information sources characterized by heterogeneous formats and structures. The paper first introduces DL_P syntax and semantics. Then, a DL_P-based approach is illustrated for inferring complex knowledge assertions from information sources characterized by heterogeneities in formats and representational structures. The thus obtained complex knowledge assertions can be exploited for constructing a user profile and for improving the quality of present Web search tools.

1 Introduction

In the last years Description Logics have been adopted as the reference formalism for inferring, representing and handling knowledge about heterogeneous databases. As an example, in [5], a Description Logic is introduced for representing databases as well as structural properties holding among database objects. In [10], the Description Logic of [5] is extended by introducing plausibility factors by which a probabilistic Description Logic, called DL_P, is obtained. DL_P is exploited in [10] for inferring complex knowledge assertions, i.e. terminological and structural relationships involving several objects belonging to heterogeneous databases. [4] shows how Description Logics can be exploited for integrating databases in the context of Data Warehousing.

Recently, the enormous diffusion of the Internet led massive amounts of data to be stored not only in traditional databases but also in semi-structured information sources, such as HTML and XML documents, and this trend seems to be confirmed for the future with the generalized adoption of the XML as the standard for data exchange, thus causing obvious and increasing problems in information access and delivery. One of the most difficult problems typically found over the Internet can be summarized in the difficulty for the user to efficiently access the information she/he needs or, in the system perspective, to deliver the right information to the right user in the right format at the right time. Such difficulties basically arise from the unavailability of formalisms and associated

M.-S. Hacid et al. (Eds.): ISMIS 2002, LNAI 2366, pp. 46–54, 2002.

inference techniques capable to allow for reasoning about source content and relationships. Description Logics, however, seem to feature sufficient semantical richness to allow designers to reason about those heterogeneous information sources.

An interesting attempt in this direction is the exploitation of a particular Description Logic coupled with Datalog-like rules in the context of Information Manifold [6], a fully implemented system that provides uniform access to an heterogeneous collection of information sources on the Web. Another interesting attempt is described in [3], where a particular Description Logic is given for representing and reasoning with Document Type Definitions of XML documents. Given a DTD, the paper shows how this can be translated in a DL knowledge base. Then, given several DTDs, by reasoning with their associated knowledge bases, the paper shows how to state several interesting relationships thereof, namely, strong and structural equivalence, inclusion, disjointness of XML document sets conformant to the given DTDs and conformance of an XML document to a given DTD. Finally, it is worthwhile to mention the application of the Description Logics in MOMIS [2], a system for the integration and query of multiple, heterogeneous information sources, containing structured and semi-structured data. Here, a particular Description Logic, called OLCD (Object Language with Complements allowing Descriptive cycles), derived from the KL-ONE family [13], is the core of the module ODB-Tools [1] which exploits OLCD along with suitable Description Logic inference techniques for *(i)* building a consistent Common Thesaurus of involved information sources, which has the role of a shared ontology for them and *(ii)* providing support for semantic optimization of queries at the global level, based on defined mapping rules and integrity constraints. Other and different exploitations of DL within the database context regard the definition of hybrid languages, notably, the CARIN language family [7]. CARIN languages are defined by merging DATALOG with DL. Concepts, as well as role predicates, are allowed in rule bodies. Recursion is also allowed. It is notable that these languages have been applied in database integration scenarios.

In this paper our aim is to continue in this attempt of verifying the possibility of exploiting Description Logics for inferring, representing and handling knowledge about information sources of different nature. In particular, we show that DL_P, with the support of a particular conceptual model called SDR-Network [12], can be used to: *(i)* represent information sources having different formats and structures; *(ii)* infer complex knowledge assertions from these sources; *(iii)* handle and exploit these inferred assertions.

In the following, we therefore first introduce the syntax and the semantics of the DL_P. Then, we illustrate the SDR-Network, a conceptual model these authors recently proposed for describing both structured and semi-structured information sources, as well as for deriving simple properties holding for them (Section 2). A set of rules is also provided that produce a DL_P representation of an SDR-Network (Section 3). Then, the paper describes the DL_P-based inference mechanism for reasoning about information structures by extracting complex knowledge patterns from them. Intuitively, these patterns are DL_P asser-