Chapter 6

Query Processing in a Mediator Based Framework for Linked Data Integration

Vânia M. P. Vidal
Federal University of Ceará, Brazil

José A. F. de Macêdo
Federal University of Ceará, Brazil

Marco A. Casanova
Pontifícia Universidade Católica do Rio de Janeiro, Brazil

Fábio Porto
Laboratório Nacional de Computação Científica, Brazil

João C. Pinheiro
Federal University of Ceará, Brazil

ABSTRACT

In this paper, the authors present a three-level mediator based framework for linked data integration. In the approach, the mediated schema is represented by a domain ontology, which provides a conceptual representation of the application. Each relevant data source is described by a source ontology, published on the Web according to the Linked Data principles. Each source ontology is rewritten as an application ontology, whose vocabulary is restricted to be a subset of the vocabulary of the domain ontology. The main contribution of the paper is an algorithm for reformulating a user query into sub-queries over the data sources. The reformulation algorithm exploits inter-ontology links to return more complete query results. The approach is illustrated by an example of a virtual store mediating access to online booksellers.

1. INTRODUCTION

The Semantic Web is attempting to provide technologies for effectively publishing, retrieving and integrating RDF data distributed over the web. We agree with (Langegger, Woss, & Blochl, 2008) that large-scale data integration is probably one of the best use cases for the Semantic Web technology. There are several aspects of the Semantic Web that make it appropriate for the integration of data from distributed and heterogeneous data sources (Wache et al., 2001). Briefly, these are: RDF, the Resource Description Framework, a simple, but powerful and extensible data model; URIs (or IRIs) used for global naming; and the possibility of reasoning based on Description Logic (Calvanese et al., 2008).
In this paper, we consider the problem of designing data integration systems (Lenzerini, 2002) when the data sources are published on the Web according to the Linked Data principles (Bizer, Heath, & Berners-Lee, 2009), which require the identification of entities with URI references that can be resolved over the HTTP protocol into RDF data that describes the identified entity. These descriptions may include RDF links pointing to other data sources. RDF links take the form of RDF triples, where the subject of the triple is an URI reference in the namespace of one data source, while the object is a URI reference in the namespace of the other. The notion of identity is an important issue in the Semantic Web. URIs guarantee that resources are uniquely identifiable resources on the Web, but they do not guarantee the uniqueness of the entities the resources refer to (Halpin, 2006). Thus, there is a need for a service that is able to find different URIs that refer to the same real-world entity.

In this paper, we propose a mediator-based framework for implementing data integration over linked data. We provide a sound and complete algorithm for reformulating a SPARQL query into a query over the (linked) data sources. The reformulation algorithm exploits inter-ontology links to return more complete query results.

This paper is organized as follows. Section 2 describes the framework proposed for linked data integration. Section 3 summarizes some basic concepts required in the paper. Section 4 presents an example that will be used throughout the paper. Section 5 discusses the query answering method adopted. Section 6 introduces a strategy for query reformulation, which is the central contribution of the paper. Section 7 lists related work. Finally, Section 8 presents the conclusions and directions for future research.

2. A FRAMEWORK FOR LINKED DATA INTEGRATION

In this section, we discuss the three-level architecture for linked data integration, which is depicted in Figure 1.

The mediated schema is represented by a domain ontology (DO), which provides a conceptual representation of the domain (a globally shared vocabulary and a set of constraints). Each relevant data source is described by a source ontology, published on the Web according to the Linked Data principles, thereby becoming part of the Web of linked data. These source ontologies are depicted in the Web of Linked Data layer in Figure 1.

The local source schemas are accessed via wrappers, like those introduced in Berners-Lee et al. (2006) which export the local data into