An Ontology-Based Framework for Heterogeneous Data Sources Integration

Vânia M. P. Vidal[1], João C. Pinheiro[1,2], Eveline R. Sacramento[1], José Antonio Fernandes de Macêdo[1], Bernadette F. Lóscio[1]

[1]Department of Computing, Universidade Federal do Ceará, Brazil
{vvidal, joaoslz, eveline, jose.macedo, bernafarias}@lia.ufc.br

[2]Department of Informatics, Instituto Federal de Educação, Ciência e Tecnologia do Maranhão, Brazil

1 The Proposed Framework

Ontologies have been extensively used to model domain-specific knowledge. The main reason for this success is due to their capability to be at the "semantic" level, away from data structures and implementation strategies. In addition, ontology formalisms have allowed certain kinds of reasoning to be automated within a reasonable time complexity. Due to ontology data independence and automated reasoning, ontologies are well suited for integrating heterogeneous databases, enabling interoperability among disparate systems, and specifying interfaces to independent, knowledge-based services.

Recent research has used ontologies for specifying the mediated schema in the context of data integration [2, 3]. An important challenge in ontology-based data integration systems is the problem of rewriting a query specified in terms of the domain ontology into queries that can be answered by the individual data sources. Reasoning is used to determine whether existing ontology concepts (describing the local data sources) are a match for the user's query and to query rewrite. Description Logics can be used to describe relationships between data sources [1] and to provide more flexible mechanisms for semantic query rewriting needed in such systems [2], Despite those approaches, little work has been done focusing on the optimization of querying processing in a ontology-based data integration system.

In this work, we propose an ontology-based framework for integration of heterogeneous data sources, comprising the benefits of ontologies to make the semantic integration; and of ontological reasoning to discard sub-queries that are not consistent and to infer additional relations between concepts. Figure 1 describes the main components of the proposed architecture. The mediated schema is represented by a domain ontology (DO), which provides a conceptual representation of the application domain (a global shared vocabulary).

Each local source schema is described by an application ontology (AO) whose vocabulary is restricted to be a subset of the vocabulary of DO. The global ontology (GO) consists of: i) a set of global axioms, which define inter-ontology properties to extract the remaining properties located in other application ontologies in order to obtain the maximum of information; and ii) the union of the application ontologies, which are a notational convenience to divide the definition of the mappings into two stages: the definition of the mediated mappings and the definition of the local mappings. The mediated mappings define the concepts and properties of the domain ontology in terms of the vocabularies of the global ontology, whereas the local mappings specify the correspondences between an application ontology and its local source schema.

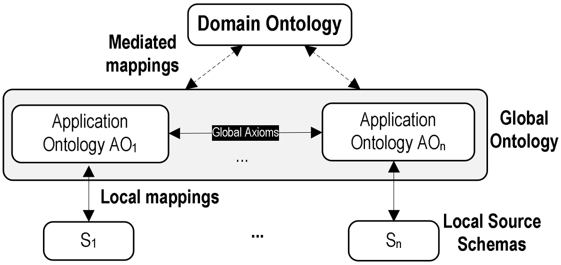

Figure 1. Ontology-based Architecture for Data Integration.

We adopt RDF/OWL to represent the domain ontology and the application ontologies. In order to represent the mediated mappings, we adopt a family of logics called Description Logics (DL). Additionally, for posing queries on the domain ontology, we use SPARQL query language. Local source schemas are accessed via wrappers exporting their data into RDF/OWL format, according to respective AO.

In our approach, the process of answering a query posed on the mediated schema consists of three steps: (i) Semantic rewriting. The user's query is decomposed into a set of elementary sub-queries over the application ontologies using the algorithm proposed in [3], and it is generated a semantic execution plan (SEP), which specifies how the results of these sub-queries are combined to obtain an intermediate result. This step is performed using the mediated mappings and the global axioms (ii) Optimization. This step consists in building a feasible and cost-effective final execution plan (FEP), considering the limitations on the access to local data sources. (iii) Evaluation. The sub-queries over the AOs are rewritten in terms of their corresponding local source schemas, with the help of the local mappings. The results of these sub-queries are returned to the mediator, where the final result is built according to the FEP. To the best of our knowledge, this is the first solution that uses an ontology as a mediated schema addressing the problem of efficient query processing on multiple data sources.

2 References

1. Calvanese, D., De Giacomo, G., Lenzerini, M., Lembo, D., Poggi, A., Rosati, R.: MASTRO-I: Efficient Integration of Relational Data through DL Ontologies. In: Proceedings of the Description Logic Workshop (DL'07), 227 – 234 (2007).
2. Lehti, P. and Fankhauser, P.: XML Data Integration with OWL: Experiences and Challenges. In: Proceedings of the Symposium on Applications and the Internet, 160-167 (2004).
3. Vidal, V.M.P., Sacramento E. R., Macedo, J. A. F., Casanova M. A.: An Ontology-Based Framework for Geographic Data Integration. In: Proceedings of Secogis (2009).