INFINT 2007 - Bertinoro Workshop on Information Integration

Data and Service Discovery and Integration on the Semantic Web

(Position Statement)

Dimitrios Skoutas
(joint work with Alkis Simitsis, Verena Kantere, and Timos Sellis)

Information integration constitutes a significant problem in data management. The main challenge arises from the structural and, especially, the semantic heterogeneity of the data sources to be integrated. The lack of precise metadata regarding the semantics of the data sources hinders the automation of the integration process. The problem is even more prominent on the Web, where the provided information is primarily intended for human browsing, and content creation and publishing takes place in a completely decentralized manner. Moreover, the advent of Web 2.0 has further blurred the distinction between content providers and content consumers, significantly contributing both to the amount of available information as well as its diversity. In this setting, the emergence of the Semantic Web raises new opportunities and challenges for the persistent problems of data and service discovery and integration. Ontologies provide a means to formally represent the knowledge for a given domain, making explicit the semantics of the involved entities, and, thereby, providing the ability to reason about them.

Our research concerns the use of Semantic Web technologies to address several important aspects regarding interoperability in heterogeneous systems. In particular, we focus on three main directions: (a) the use of ontologies for designing ETL processes, (b) a semantic similarity measure for identifying relevant peers and assessing the quality of query rewritings in P2P database management systems (PDMS), and (c) the discovery and composition of Semantic Web services.

ETL. In previous work, we have proposed an ontology-based approach for facilitating the conceptual design of ETL processes. Each source or target schema is represented by means of a graph, called *datastore graph*. The construction of an appropriate application ontology is described, which is specified using the Web Ontology Language (OWL). The ontology provides the ability for modelling: (a) the concepts of the domain, (b) the relationships between those concepts, (c) the attributes characterizing each concept, and (d) the different representation formats and (ranges of) values for each attribute. A graph representation, called *ontology graph*, is also adopted for the ontology. The source and target schemas are then annotated, by establishing correspondences between elements of the datastore graph and the ontology graph. Based on the ontology and the mappings, a reasoning technique is applied to infer correspondences and conflicts between the datastores. This allows us to identify relevant sources for populating a given target, as well as required conceptual operations to be included in the ETL process, so that data is appropriately transformed from the sources to the target. Still, more in-depth analysis and evaluation is required, the latest, however, being encumbered by the lack of an established methodology and testbeds. Another issue is to support the evolution of the ETL scenarios, when changes in the underlying data sources occur.

PDMS. Concerning data sharing in peer-to-peer databases, the available rewriting algorithms for structured data target the classic data integration problem and consider only queries that can be completely rewritten to the target schema under a set of mappings. Still, such approaches are not enough for an environment where peers seek, and are satisfied with, information semantically similar, but not necessarily identical, to their requests. Hence, there is a necessity for investigating the notion of semantic similarity of peer schemas, and, furthermore, of peer queries with their rewritten versions. Using such similarity criteria, users can identify peers sharing similar interests to theirs, and decide, for each specific query they pose, which peers can rewrite it better. To this end, an initial investigation has been presented, based on the use of a shared ontology to semantically annotate the schema of each peer, making explicit the type of information provided by it. Peers share structured data through the use of schema mappings. When a query is forwarded and is rewritten among peers, according to the corresponding pairwise mappings, it is often degraded. This is because (a) a portion of the attributes of the original query may not be rewritten, or may be rewritten poorly, and (b) conditions existing in the original query may be lost, or additional ones may be inserted, due to the nature of the mappings. A semantic similarity measure is proposed to assess the quality of the rewriting, taking into account (a) the semantics of the peer schemas, as specified by the annotations, (b) the pairwise mappings between the peers, and (c) the particular queries posed. An issue to be addressed is to investigate the notion of synopsis of similarity values, so as to guide the propagation of queries to the most relevant peers in the overlay. Another challenge is to adapt this method to a social network application and to evaluate its effectiveness in that domain.

Web Services. The use of ontologies has also been proposed for adding semantics to the descriptions of Web services, in order to allow software agents to process and reason about these descriptions, achieving a high degree of automation for tasks such as service discovery and composition. Matchmaking for Semantic Web services is based on the use of logic inference to check for equivalence or subsumption relationships between ontology classes. In particular, existing approaches use a discrete scale to specify the degree of match between a service request and a service description: *exact*, *plug-in*, *subsume*, *intersection* and *disjoint*. However, this coarse-grained categorization is inadequate, when dealing with a large number of candidate services. Instead, a ranking of the available services is required, so that the search can be restricted to the results having the highest scores. In previous work, we have proposed an approach for ranking semantically annotated Web services with respect to a service request. Recall and precision, two well-known evaluation measures adapted from the area of Information Retrieval, are used to express the degree of match between the service description and the service request. The matching function is asymmetric, allowing to distinguish whether the service capabilities are a superset or a subset of the ones being requested. The degree of match is assessed based on matching input and output parameters, as well as preconditions and effects, according to the semantic information conveyed in the domain ontology. The results are specified in a continuous scale, while at the same time maintaining a correspondence to the types of match mentioned previously. Future plans include the extension of the matching mechanism to incorporate QoS aspects, as well as its application to the problem of service composition. Furthermore, we plan to investigate the use of the proposed semantic similarity measure for constructing an efficient indexing scheme for Semantic Web service discovery in structured peer-to-peer networks.