Towards Distributed RDF Querying

Richard Vdovjak\textsuperscript{1}, Geert-Jan Houben\textsuperscript{1},
Heiner Stuckenschmidt\textsuperscript{2}
\textsuperscript{1} Technische Universiteit Eindhoven
\textsuperscript{2} Vrije Universiteit Amsterdam

Abstract

The need for handling multiple sources of knowledge and information is very apparent in the context of Semantic Web applications. First of all, we have the duality of schema and information content where multiple information sources can adhere to the same schema. Further, the reuse, extension and combination of multiple schemata is considered to be common practice on the Semantic Web. Despite the inherently distributed nature of the Semantic Web, most of the current RDF infrastructures store information locally as a single knowledge repository, i.e. RDF models from remote sources are replicated and merged into a single model. Distribution is virtually retained through the use of namespaces to distinguish between different models. We argue that many interesting applications on the Semantic Web would benefit from or even require an RDF infrastructure that supports real distribution of information sources that can be accessed from a single point.

In this talk we address the problem of how to specify the integration of (possibly heterogeneous) information sources on the Semantic Web in order to facilitate a uniform access to their distributed data. Such transparent data access is desirable in many areas. As an example, we show its benefits within the Hera context of Web information system design. We propose an Integration Model formalism which is able to deal with many semantic heterogeneities that frequently occur among sources on the Semantic Web. We also address the issues of distributed query processing and optimization in such a framework. Our approach is based on an existing centralized RDF storage and retrieval system, the architecture of which was extended towards a distributed setting. From this architecture we derive a set of requirements for the design of access methods. We present techniques for distributed RDF query processing, together with an index structure for locating information in different sources. Further, we map the problem of RDF (path) querying to that of join ordering, and propose the use of optimization techniques known from the database technology to improve the performance of our system.