Dissertation

„P2P based RDF Querying and Reasoning for Grid Resource Description and Matching“

Felix Heine

Abstract

In this thesis, we look at the problem how resources in large, heterogeneous Grids can be discovered. In world-sized Grids, two aspects of the **resource discovery** problem become especially important: **heterogeneity** and **size**.

Heterogeneity means that the types of resources included in the Grid are highly diverse. Additionally to traditional resources like high performance clusters and storage devices, any kind of service including arbitrary applications and expensive physical instruments are treated as resources. No single standard can encompass any resource to be described. As soon as there are multiple standards, additional knowledge is needed to mediate between these standards.

Size means that a scalable solution to the resource discovery problem is needed. Although there are numerous reasoning systems, they typically assume that all knowledge is collected at a single system, which is infeasible for arbitrary large collections of information.

We present a system that contributes the initial steps to the solution of the described problem. Although Grid computing is the motivating application, the scope of this thesis is larger. Its goal is to provide **a scalable approach to combine and query large-scale collections of machine-readable information**.

The main elements of the thesis are the description of the system architecture based on a structured p2p network, a dissemination algorithm that places information on well-defined nodes, a reasoning mechanism that derives new knowledge from the existing, combining information which originates from different nodes, and two query evaluation strategies.

The first evaluation strategy aims to extract all matches for a given query. We describe various strategies to minimize the network load. However, in case of queries with large result sets, an exhaustive evaluation is infeasible. Thus we present a second strategy targeting queries with a huge number of results that retrieves only a restricted number of results according to some sorting criterion. We use caching and look ahead strategies to make the algorithm efficient.

We have implemented the system prototypically. Using this implementation, we perform various experiments both on a simulation base and using real test runs to show the efficiency of the system.

Schlüsselwörter

Grids, Resource Description, Resource Matching, Semantic Grid, Information Integration, Semantic Web, RDF, RDFS, Reasoning, SPARQL, Query Processing, Top k Query Processing, Scalability, Data Organization, P2P, DHT, Structured Overlay Networks