Chapter 23
Hypermedia-Based Discovery for Source Selection Using Low-Cost Linked Data Interfaces

Miel Vander Sande
Data Science Lab, Ghent University - iMinds, Belgium

Ruben Verborgh
Data Science Lab, Ghent University - iMinds, Belgium

Anastasia Dimou
Data Science Lab, Ghent University - iMinds, Belgium

Pieter Colpaert
Data Science Lab, Ghent University - iMinds, Belgium

Erik Mannens
Data Science Lab, Ghent University - iMinds, Belgium

ABSTRACT

Evaluating federated Linked Data queries requires consulting multiple sources on the Web. Before a client can execute queries, it must discover data sources, and determine which ones are relevant. Federated query execution research focuses on the actual execution, while data source discovery is often marginally discussed—even though it has a strong impact on selecting sources that contribute to the query results. Therefore, the authors introduce a discovery approach for Linked Data interfaces based on hypermedia links and controls, and apply it to federated query execution with Triple Pattern Fragments. In addition, the authors identify quantitative metrics to evaluate this discovery approach. This article describes generic evaluation measures and results for their concrete approach. With low-cost data summaries as seed, interfaces to eight large real-world datasets can discover each other within 7 minutes. Hypermedia-based client-side querying shows a promising gain of up to 50% in execution time, but demands algorithms that visit a higher number of interfaces to improve result completeness.

DOI: 10.4018/978-1-5225-5191-1.ch023

INTRODUCTION

The Web is a fully distributed system—and thus so is the Web of Data. Within this enormous collection, each data source specializes in its very own part of the truth. Some of them, like DBpedia1, contain essential facts about a broad range of subjects; others, like Drugbank2, offer a comprehensive corpus of triples about highly select topics. As a result, in order to answer any non-trivial query over the Web of Data, we likely need to consult multiple data sources. The need for such federated queries intensifies as the Linked Open Data cloud is trending toward a more decentralized graph structure, with additional linking hubs besides DBpedia arising (Schmachtenberg et al., 2014). Federation is thus necessary to achieve the Web of Data vision (Heath & Bizer, 2011): a global, machine-understandable dataspace with web-scale integration and interoperability.

In literature, the story of federated query evaluation is typically told from source selection onwards: given a fixed set of available data sources, a client determines which of these are necessary to obtain results. After that, the actual query processing against the selected sources happens. However, before any of this can take place, candidate data sources need to be located first. This process preceding source selection has hardly received rigorous scientific study so far. In general, discovery is the process of finding available Linked Data sources that are relevant to a certain task, for specific definitions of "relevance" and "task". Although the description of dataset or endpoint characteristics has been covered, the act of finding, accessing, and processing such documents is still in its infancy. With the emerging Web Of Data, studying autonomous Linked Data discovery becomes a need, with a special focus on the impact on client-side tasks such as querying. For federated query execution in particular, discovery can assist in a more complete selection of accessed data sources.

Therefore, this article studies the impact of Linked Data interface discovery on federated querying. We consider any that provides client access to Linked Data sources. In total, we present three contributions.

First, we propose a discovery technique, which leverages hypermedia between Linked Data interfaces. Hypermedia allows such interfaces to function similarly to a webpage, providing the user with guidance on what type of content they can retrieve, or what actions they can perform, as well as the appropriate links to do so. Since the beginning of the Web, this has been the crucial aspect to the Web's scalability. Existing discovery works have greatly progressed in closed, custom p2p networks using custom discovery protocols, or centralized repositories that crawl metadata from different sources. However, with a scale-free network at our disposal, little of its benefits have been exploited for Linked Data querying. The novelty of our approach lies in strictly reusing hypermedia and Linked Data principles to a) discover one another, aided by links in a dataset; and b) inform the client at run-time about their discoveries through hypermedia. Furthermore, clients and servers distribute the processing cost fairly, resulting in a sustainable and scalable solution.

Second, to appropriately evaluate discovery approaches, we introduce a methodology to quantify its parameters. This includes metrics to express the functional and non-functional characteristics of one discovery approach relative to others.

Third, we implement and evaluate the approach against the lightweight Triple Pattern Fragments interface (Verborgh et al., 2014; 2016), and measure to what extent our discovery method facilitates source selection in federated query execution. We intend to enable querying multiple sources on the client while obtaining far less information than heuristics or dataset profiles.

The remainder of this paper is structured as follows. We first list a number of research questions with corresponding hypotheses and discuss related work. Then, we propose the metrics for evaluating