Semantic and Distributed Entity Search in the Web of Data

Robert Neumayer
Norwegian University of Science and Technology
Trondheim, Norway
neumayer@idi.ntnu.no

Abstract

Both the growth and ubiquitous character of the Internet have had a profound effect on how we access and consume data and information. More recently, the Semantic Web, an extension of the current Web has come increasingly relevant due to its widespread adoption.

The Web of Data (WoD) is an extension of the current web, where not only documents are interlinked by means of hyperlinks but also data in terms of predicates. Specifically, it describes objects, entities or "things" in terms of their attributes and their relationships, using RDF data (and often is used equivalently to Linked Data). Given its growth, there is a strong need for making this wealth of knowledge accessible by keyword search (the de-facto standard paradigm for accessing information online).

The goal of this thesis is to provide new techniques for accessing this data, i.e., to leverage its full potential to end users. We therefore address the following four main issues: a) how can the Web of Data be searched by means of keyword search?, b) what sets apart search in the WoD from traditional web search?, c) how can these elements be used in a sound and effective way?, and d) How can the techniques be adapted to a distributed environment?

To this end, we develop techniques for effectively searching WoD sources. We build upon and formalise existing entity modelling approaches within a generative language modelling framework, and compare them experimentally using standard test collections. We show that these models outperform the current state-of-the-art in terms of retrieval effectiveness, however, this is done at the cost of abandoning a large part of the semantics behind the data. We propose a novel entity model capable of preserving the semantics associated with entities, without sacrificing retrieval effectiveness. We further show how these approaches can be applied in the distributed context, both with low (federated search) and high numbers (Peer-to-peer or P2P) of independent repositories, collections, or nodes.

The main contributions are as follows:

- We develop a hybrid approach to search in the Web of Data, using elements from traditional information retrieval and structured retrieval alike.
- We formalise our approaches in a language model setting.
- We discuss and analyse based on our empirical evaluation and provide insights into the entity search problem.