Exploiting Linked Data in Exploratory Search

Pavlos Fafalios
Institute of Computer Science, FORTH-ICS, Greece, and
Computer Science Department, University of Crete, Greece
fafalios@ics.forth.gr

Abstract

In recent years we have witnessed an explosion in publishing data on the Web, mostly in the form of Linked Data. An important question is how typical users, who mainly use keyword search queries, can access and exploit this constantly increasing body of knowledge. Although existing interaction paradigms in Semantic Search hide their complexity behind easy-to-use interfaces, they have not managed to cover common search needs. At the same time, according to several studies, a large number of search tasks are of exploratory nature. However, in such tasks the traditional "ranked list" approach for interacting with the retrieved results is often inadequate.

The objective of this thesis is to enable effective exploratory search services which can bridge the gap between the classic responses of non-semantic search systems (e.g., Professional Search Systems, Web Search Engines) and semantic information expressed in the form of Linked Open Data (LOD). Towards this direction, we introduce an approach in which named entities (like names of persons, locations, chemical substances, etc.) are exploited as the glue for automatically connecting documents (search results) with data and knowledge. We study an approach where this entity-based integration is performed at real-time, without any human intervention and without the need of prebuilt indexes. This allows the provision of "fresh" information, the easy configuration of this functionality according to the needs of the underlying search application, as well as its easy exploitation by existing search systems.

The provision of the aforementioned functionality is challenging. At first, the LOD that are available on the Web are big, are distributed in many knowledge bases, are increased and updated continuously, and also cover many domains. Consequently, there is the need of an interoperability model that will allow the specification of the entities of interest as well as of the related and useful semantic data. In addition, the number of extractable entities from the search results can be very high and the same is true for the amount of semantic information that can be retrieved from the LOD for these entities (i.e., the number of their attributes and of their associations with other entities). Thus, there is also the need of methods that can estimate the important (for the search context) entities, attributes and associations.

To cope with the above challenges, this thesis proposes a semantic analysis process in which the search results are connected with data and knowledge at real-time without any human intervention. For describing the entities of interest, as well as the related (and useful for the application context) semantic information, we propose a generic model for configuring a Named Entity Extraction (NEE) system, while for specifying the semantics of this model, we introduce an RDF/S vocabulary, called "Open NEE Configuration Model", which allows

a NEE system to describe (and publish as LOD) its entity-mining capabilities. To enable associating the result of a NEE process with an applied configuration, we propose an extension of the Open Annotation Data Model which also allows publishing the annotation results as LOD. To examine the feasibility of this model, we developed the system X-Link which, contrary to existing NEE systems, allows its easy configuration by exploiting one or more semantic Knowledge Bases. To identify the important semantic information related to the search results, we introduce and study a ranking method that is based on the Random Walk model and which exploits the extracted entities and their connectivity. The exploitation of the selected semantic information is achieved either through the visualization of the related semantic graph and/or in the context of a faceted interaction model that allows the user to gradually restrict the search space. Besides, this thesis studied the exploitation of such graphs for re-ranking the list of retrieved results aiming to promote relevant but low-ranked hits.

The dissertation reports extensive evaluation results of the proposed functionalities and methods. Regarding the system X-Link, a task-based evaluation with users showed its ease of configuration, while a case study illustrated the efficiency of the supported operations. The comparative evaluation of the proposed probabilistic scheme for ranking entities and semantic data showed that the proposed approach is more effective compared to other ranking approaches (producing a more than 20% better ranking). Regarding the presentation of the important entities (and of their associations), the conducted survey in a marine-related search context demonstrated that the majority of participants (more than 70%) prefer to see a graph representation of entities related to the retrieved results regardless of the type of the submitted query. The evaluation of the proposed probabilistic algorithm for re-ranking the retrieved search results (using TREC datasets related to the medical domain) showed that this approach can notably improve the list of results by promoting relevant hits in higher positions. Finally, the implementation and the experimental results of the proposed search process demonstrated its feasibility and efficiency, and also enabled us to reveal its limitations.

Supervisor: Yannis Tzitzikas (Computer Science Department, University of Crete, Greece)
Available at: `http://users.ics.forth.gr/~fafalios/Dissertation_Fafalios.pdf`