Entity Linking and Retrieval for Semantic Search

Edgar Meij
Yahoo! Research
Barcelona, Spain
emeij@yahoo-inc.com

Krisztian Balog
University of Stavanger
Stavanger, Norway
krisztian.balog@uis.no

Daan Odijk
ISLA, University of Amsterdam
Amsterdam, The Netherlands
d.odijk@uva.nl

ABSTRACT

More and more search engine users are expecting direct answers to their information needs, rather than links to documents. Semantic search and its recent applications enabled search engines to organize their wealth of information around entities. Entity linking and retrieval provide the building stones for organizing the web of entities. This tutorial aims to cover all facets of semantic search from a unified point of view and connect real-world applications with results from scientific publications. We provide a comprehensive overview of entity linking and retrieval in the context of semantic search and thoroughly explore techniques for query understanding, entity-based retrieval and ranking on unstructured text, structured knowledge repositories, and a mixture of these. We point out the connections between published approaches and applications, and provide hands-on examples on real-world use cases and datasets.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software

Keywords

Entity linking, entity retrieval, semantic search

1. OVERVIEW

The explosive increase in the amount of unstructured textual data being produced calls for advanced methodologies for making sense of this data. Recent advances have enabled a precise manner of analysis, where phrases—consisting of a single term or sequence of terms—are automatically linked to entries in a knowledge base. This process is commonly known as *entity linking*. Entity linking facilitates advanced forms of searching and browsing in various domains and contexts. For example, it can be used to anchor textual resources in background knowledge. In search engines, linking queries to entities to improve the user experience is becoming increasingly prevalent. More and more, users want to find the actual entities that satisfy their information need, rather than merely the documents that mention them; a process known as *entity retrieval*.

Entities are a key enabling component for semantic search, as many information needs can be answered by returning a list of entities, their properties, and/or their relations. They can be used to enrich the search result page to enable direct answers or serendipitous results, and are able to bridge the gap between unstructured and structured data.

The goal of this tutorial is to cover all facets of semantic search from a unified point of view and connect real-world applications with results from scientific publications. Its main highlights are that we: (i) provide a comprehensive overview of entity linking and retrieval in the context of semantic search, (ii) thoroughly explore techniques for query understanding, entity-based retrieval and ranking on unstructured text, structured knowledge repositories, and a mixture of these, and (iii) point out the connections between published approaches and applications, and provide hands-on examples on real-world use cases and datasets.

The half-day tutorial is structured into five parts. Part I sets the scene by introducing the data landscape. We make a case for entity retrieval that is built on the observation that web search engine users prefer to express information needs using short keyword queries and that many of these revolve around entities. Part II zooms in on the congruent relationship between textual and structural evidence for entities. These come in two main flavors: (i) entity linking and (ii) knowledge base population/acceleration, i.e., given a massive text corpus, populate a knowledge base with new entities, relations, and attributes. Part III starts with an overview of traditional methods for keyword-based search in the IR, DB, and semantic web communities. Then, we discuss methods that combine these aspects and where query understanding may involve some inferencing. Part IV continues with settings where the aim is to provide the user with an improved overall search experience, e.g. by providing an enriched SERP which lists recent match results when searching for a football team or displaying restaurants on a map. Alternatively, the standard document-based result list can be augmented with entity-oriented components for serendipitous discovery. Finally, Part V concludes the tutorial with an overview and an outlook of future challenges. Each block discusses a particular topic and is be followed by hands-on exercises, illustrating the theory presented before.

The slides, bibliography, and resources of the current and previous editions can be found at `http://bit.ly/ELR-slides`.

WSDM'14, February 24–28, 2014, New York, New York, USA.
ACM 978-1-4503-2351-2/14/02.
http://dx.doi.org/10.1145/2556195.2556201.