Entity Linking and Retrieval

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

This full-day tutorial presents a comprehensive introduction to entity linking and retrieval. Part I provides a detailed overview of entity linking: identifying and disambiguating entity occurrences in unstructured text. Part II focuses on entity retrieval, by first considering scenarios where explicit representations of entities are available, and then moving to a setting where evidence needs to be collected and aggregated from multiple documents or even collections, thereby combining techniques from both entity linking and entity retrieval. Part III concludes the tutorial with an overview and hands-on comparative analysis of applications and publicly available toolkits and web services.

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 in all kinds of domains 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. It can be used, for instance, to anchor textual resources in background knowledge; authors or readers of a piece of text may find entity links to supply useful pointers. Another application can be found in search engines, where 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*.

It is common to consider entities from a general-purpose knowledge base such as Wikipedia or Freebase, since these provide sufficient coverage for most tasks and applications. Wikipedia is therefore a typical target for entity linking and also a fertile ground for entity retrieval; its rich structure (including wikilinks, categories, and infoboxes) informs entity linking algorithms and enables advancing over plain document retrieval. Approaches for linking and retrieving entities are not Wikipedia-specific, however. Recent developments in the Web of Data enable the use of domain or task-specific entities. Alternatively, legacy or corporate knowledge bases can be used to provide entities. Entity linking and retrieval is also gaining popularity in the public domain, as witnessed by Wolfram Alpha, the Google Knowledge Graph, digital personal assistants such as Siri and Google Now, and various vertical search engines focusing on particular entity types.

In this full-day tutorial we present a comprehensive introduction to entity linking and retrieval. Part I of the tutorial provides a detailed overview of entity linking. We introduce the fundamental concepts and principles, address the identification and disambiguation of entity occurrences in unstructured text, and detail state-of-the-art algorithms including unsupervised solutions, graph-based methods, and feature-based approaches in a machine learning setting. We continue with applications of entity linking for IR and conclude this part with discussing evaluation methodology.

Part II focuses on entity retrieval and begins with a study of scenarios where explicit representations of entities are available in the form of, e.g., Wikipedia articles or RDF triples. We continue in a setting with more complex queries, requiring evidence to be collected and aggregated from unstructured textual data—with the potential help of some structured data. Such queries require a combination of techniques from entity linking and entity retrieval. Throughout Part II, two main families of models are discussed: generative language models and discriminative feature-based models. Both the entity linking and entity retrieval parts are anchored in recent evaluation efforts conducted at benchmarking campaigns such as INEX, TAC, and TREC. We introduce test collections, tasks, evaluation methodology, and experimental results from these evaluation initiatives.

Finally, a number of publicly available toolkits and web services for entity linking and entity retrieval exist. The last part of the tutorial will give an overview and comparative analysis of these, followed by a hands-on session where they will be evaluated in various settings.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
SIGIR'13, July 28–August 1, 2013, Dublin, Ireland.
ACM 978-1-4503-2034-4/13/07.