Exploiting Entities for Query Expansion

Wladmir Cardoso Brandão
Computer Science Department
Universidade Federal de Minas Gerais
Av. Antônio Carlos, 6627 - ICEx, 4010
Belo Horizonte, MG, Brasil
wladmir@dcc.ufmg.br

April, 2014

Abstract

A substantial fraction of web search queries contain references to entities, such as persons, organizations, and locations. This significant presence of named entities in queries provides an opportunity for web search engines to improve their understanding of the user's information need.

In this thesis, we investigate the entity-oriented query expansion process. Particularly, we propose two novel and effective query expansion approaches that exploit semantic sources of evidence to devise discriminative term features, and machine learning techniques to effectively combine these features in order to rank candidate expansion terms. As a result, not only do we select effective expansion terms, but we also weigh these terms according to their predicted effectiveness. In addition, since our query expansion approaches consider Wikipedia infoboxes as a source of candidate expansion terms, a frequent obstacle is that only about 20% of Wikipedia articles have an infobox. To overcome this problem we propose WAVE, a self-supervised approach to autonomously generate infoboxes for Wikipedia articles.

First, we propose UQEE, an unsupervised entity-oriented query expansion approach, which effectively selects expansion terms using taxonomic features devised by the semantic structure implicitly provided by infobox templates. We show that query expansion using infoboxes presents a better trade-off between retrieval performance and query latency. Moreover, we demonstrate that the automatically generated infoboxes provided by WAVE are as effective as manually generated infoboxes for query expansion. Lastly, we propose L2EE, a learning to rank approach for entity-oriented query expansion, which considers semantic evidence encoded in the content of Wikipedia article fields, and automatically labels training examples proportionally to their observed retrieval effectiveness.

Experiments on three TREC web test collections attest the effectiveness of L2EE, with significant gains compared to UQEE and state-of-the-art pseudo-relevance feedback and entity-oriented pseudo-relevance feedback approaches.

Available on-line at `http://hdl.handle.net/1843/ESBF-9GMJW2`.