Entity Centric Information Retrieval

Xitong Liu
University of Delaware, Newark, DE, USA
xtliu@udel.edu

Abstract

In the past decade, the prosperity of the World Wide Web has led to fast explosion of information, and there is a long-standing demand on how to access such a huge volume of information effectively and efficiently. Information Retrieval (IR) aims to tackle the challenge by exploring approaches to obtain relevant information items (e.g., documents) relevant to a given information need (e.g., query) from a huge collection of textual data (e.g., the Web). Named entity (e.g., person, location, product, event, organization) is a type of term compound widely existing in textual data. Recent advances in Information Extraction make it possible to extract entities from large volume of free text efficiently, and the research community are actively exploring whether entities would contribute to the retrieval effectiveness.

In this thesis, we investigate how to leverage entities to improve retrieval in several directions. We start with finding entities with certain semantic relation, which aims at retrieving entities and their associated attributes to meet user's information need directly. This is different from traditional search paradigm in which only documents are retrieved. Entity retrieval is performed by first retrieving a list of documents and then extracting entities from those documents. We propose a novel probabilistic framework which leverages supporting documents as bridge to model the relevance between query and entities and rank entities accordingly.

On the other side, we also explore how to leverage entities to improve effectiveness of ad hoc document retrieval in two directions. The first direction is entity-centric query expansion. We find related entities of query, and perform query expansion using the names and relations of related entities. Significant improvements over several state-of-the-art feedback models could be observed on multiple data collections. Besides, we explore another direction: entity-centric relevance modeling. We propose a novel retrieval approach, i.e., Latent Entity Space (LES), which models the relevance by leveraging entity profiles to represent semantic content of documents and queries. Experimental results over several TREC collections show that LES is effective on capturing latent semantic content and can significantly improve the search accuracy of several state-of-the-art retrieval models for entity-bearing queries.

This thesis presents a series of research efforts on entity centric information retrieval in several directions, and reveals promising potential of entities on improving the retrieval effectiveness. With the fast curation of high-quality knowledge base, more information about entities could be easily accessed and integrated into retrieval models. We hope our work could serve as guideline for future work on leveraging entities to improve information retrieval in more applications.

Supervisor: Hui Fang

Available at: `http://xtliu.com/thesis.pdf`