Exploring Topic Structure: Coherence, Diversity and Relatedness

Jiyin He
University of Amsterdam, the Netherlands
jiyinhe@gmail.com

Abstract

The use of topical information has long been studied in the context of information retrieval. For example, grouping search results into topical categories enables more effective information presentation to users, while grouping documents in a collection can lead to efficient information access. We define a topic as the main theme or subject contained in a (set of) document(s). While topics provide information about the subjects contained in a document, the structure of topics provides information such as the degree to which a set of documents is focused on certain topic (or set of topics), topical diversity among documents, and semantic relatedness of topics.

The work of this thesis focuses on modeling the structure of topics present in a (set of) document(s), with the goal of effectively using it in information retrieval. In particular, we consider a number of IR tasks where the notion of relevance is beyond "aboutness" and topic structure plays an important role in satisfying users' information need. The following research themes are addressed, in three parts: (1) topic coherence, (2) diversity and the cluster hypothesis, and (3) relating topics present in different representations.

With respect to the first research theme, we develop a coherence measure that effectively captures topical coherence of a set of documents. The proposed measure is applied to two IR tasks, namely, blog feed retrieval and query performance prediction.

In the second part of the thesis, we explore the impact of topic structure on effectively presenting retrieval results, with a focus on the scenario of result diversification. We re-visit the cluster hypothesis with respect to ambiguous or multi-faceted queries and investigate the effectiveness of query-specific clustering in result diversification.

Topics can be represented in different ways, e.g., using clusters, using definitions from a thesaurus, using statistics of term frequencies, etc. In the third part of the thesis, we study the problem of relating topics represented in different forms within the context of automatic link generation. We identify a set of significant terms from a source text, link those terms to their corresponding entries in a knowledge base in such a way that the source text is annotated with background information available in the knowledge base. We conduct a case study in automatically generating links from narrative radiology reports to Wikipedia. Such links are expected to help users understand the medical terminology and thereby increase the value of the reports. Here we evaluate state-of-the-art link generation systems and propose an approach that improves over state-of-the-art systems on radiology data.

Available online at `http://dare.uva.nl/record/377895`.