Exploring topic structure: Coherence, diversity and relatedness
He, J.

Citation for published version (APA):
He, J. (2011). Exploring topic structure: Coherence, diversity and relatedness

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
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: (1) Topic coherence; here we develop a coherence score that effectively captures topical coherence of a set of documents. The proposed score is applied to two IR tasks, namely, blog feed retrieval and query performance prediction. (2) Diversity and the cluster hypothesis, where we investigate the relation between diversity, relevance and the cluster hypothesis. 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. (3) Relating topics present in different representations. Topics can be represented in different ways, e.g., using clusters, using definitions from a thesaurus, using statistics of term frequencies, etc. 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.