Topic Structure for Information Retrieval

Jiyin He
ISLA, University of Amsterdam
Science Park 107, 1098 XG Amsterdam
j.he@uva.nl

ABSTRACT

In my research, I propose a coherence measure, with the goal of discovering and using topic structures within and between documents, of which I explore its extensions and applications in information retrieval.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Retrieval Models*

General Terms

Algorithms, Theory, Experimentation, Measurement

Keywords

Topic structure, Semantic relatedness

The use of topical information in information retrieval has long been studied. Topics in a given set of text units can be obtained by clustering semantically similar texts, or modeled implicitly with "latent semantic" methods. In my research, I propose to analyze and use the structure of the topics, i.e., the allocation of the topic-clusters in the semantic space. While the topics provide information about the semantic relatedness among texts, the structure of topics provides additional information such as the semantic relatedness among topics, how well the texts are focused on the topics, etc. In this paper, I discuss three issues related to this overall aim: (1) measuring topic structure, (2) features used for calculating semantic relatedness, and (3) applications of topic structure information in retrieval tasks.

For measuring topic structure, I propose a coherence measure [1–3] which measures the clustering structures of a given set of text units. It measures the topic structure, or clustering structure in an implicitly way, i.e., without explicitly conducting the clustering procedure, rather, it compares the distribution of the similarity scores among the given set of text units to that of a set of text units randomly drawn from the background collection. Instead of presenting the exact number of clusters, the coherence score tells us whether a set of text units is a tight cluster, a loose cluster with many sub-clusters or some randomly collected text units.

To evaluate the efficacy of the measure, I identify two ways. In our experiments, we constructed document sets that contain different number of clusters, and evaluated the coherence measure against this ground truth. The results show that the coherence score can reflect the varying number of clusters within the dataset. On top of that, the applications of coherence score so far [1–3] show that it can effectively capture the topic structure. One of the related issues is the similarity measures and the features used to represent the text units. Currently, all my experiments heuristically use the cosine similarity and lexical features, i.e., statistics of terms. In order to gain more insight of the measure as well as to explore its potential applications, I plan to experiment with different similarity measures and explore more possible features such as non-lexical features.

As a next step, my focus is on using the coherence score within various applications in IR. I focus on three types of topic structure in the context of information retrieval, viz. query topic structure, inter-document topic structure, and intra-document topic structure. Query topic structure refers to the topic structure present in the text units associated with the query, e.g., the documents retrieved by firing the query at a target collection, the documents that contain the query word, etc. In my work, I have shown that a coherence-based measure can be used for predicting the query ambiguity [2]. A further step would be attempting to use this information to help query modeling. Inter-document topic structure refers to the topic structure present among the documents in the target collection, which has wide application in cluster/topic-based retrieval models. One of the issues is to select the optimal clusters with respect to the query, where the coherence score is potentially useful, given that it reflects the tightness, i.e., the quality of clusters. Intra-document topic structure refers to the topic structure present in a single document in the target collection. In [3] we explored the use of coherence score for topic distillation in the setting of blog retrieval, which is an example task where intra-document topic structure is useful. A potential step might be toward passage-based document retrieval, where the intra-document topic structure can be measured at the passage level, which in turn can serve as an indicator of the degree to which a document is relevant to a given query.

References

[1] J. He and M. Larson and M. de Rijke. On the Topical Structure of the Relevance Feedback Set. In *WIR '08*, 2008.
[2] J. He and M. Larson and M. de Rijke. Using Coherence-Based Measures to Predict Query Difficulty. In *ECIR'08*, 2008
[3] J. He and M. Larson and M. de Rijke. Blogger, stick to your story: modeling topical noise in blogs with coherence measures. In *AND'08*, 2008

SIGIR'09, July 19–23, 2009, Boston, Massachusetts, USA.
ACM 978-1-60558-483-6/09/07.