Towards a Combined Model for Search and Navigation of Annotated Documents

Edgar Meij
ISLA
University of Amsterdam
Amsterdam, The Netherlands
emeij@science.uva.nl

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Language parsing and understanding*

General Terms

Algorithms, Theory, Experimentation, Measurement

Keywords

Language Models, Document Annotations, Semantic Relatedness

Documents whose textual content is complemented with annotations of one kind or another are ubiquitous. Examples include biomedical documents (annotated with MeSH terms) and news articles (annotated with IPTC terms). Such annotations—or *concepts*—have typically been used for query expansion, to suggest alternative or related query formulations, and to facilitate browsing of the document collection. In recent years, we have seen two important developments in this area: (i) a renewed interest in the knowledge sources underlying the annotations, mainly inspired by semantic web initiatives and (ii) the creation of social annotations, as part of web 2.0 developments. These developments motivate a renewed interest in models and methods for accessing annotated documents.

The theme of my proposed research is to capture two aspects in a single, unified model: retrieval and navigation. Given a query, this entails using both term-based and concept-based evidence to locate relevant information (retrieval) and suggesting useful browsing suggestions (navigation). I imagine this to be a "two-way" process, i.e., the user can browse the document collection using concepts and the relations between concepts, but she can also navigate the knowledge structure using the (vocabulary) terms from the documents. Such information seeking behavior is witnessed in an increasing number of applications and domains (e.g., suggesting related tags in Bibsonomy or Flickr), providing a solid motivation for my research agenda.

In order to accomplish this unification, I will first need to address three separate, but intertwined issues. First, a way of "bridging the gap" between concepts and (vocabulary) terms is needed, since concepts are not directly observable. Second, relations between concepts need to be modeled in some way. Finally, the concepts and relations thus modeled should be integrated in the information seeking process, thereby improving both retrieval and navigation.

So far, I have formulated concept modeling as a form of text classification, by representing concepts as distributions over vocabulary terms. In the context of a digital library setting, I have shown that integrating conceptual knowledge in this way can be beneficial both to retrieval performance as well as to facilitate navigation [1]. More recently, I have taken these experiments a step further by creating *parsimonious* concept models [2, 3]. In these experiments, the integration of concepts in the query model estimations is able to deliver significantly better results, both compared to a query likelihood run as well as to a run based on relevance models.

To determine the strength of relations between concepts, I have looked at using the divergence between concept models [4]. The estimations are based on differences in language use as measured by computing the cross-entropy reduction between concept models. Experimental results show that this approach is able to outperform both path-based as well as information content-based methods on two separate test sets. While this approach measures the similarity between concepts, it does not explicitly take a relation type into consideration. Thus, any explicit link structure present in the used knowledge structure disappears. Whether this is a reasonable assumption for my work is still unclear and something I intend to find an answer to.

In future work, I would also like to address the question how the retrieval-oriented models I have introduced so far may be used to further aid navigation. To some extent, I have already used the TREC Genomics test collections for the evaluation of the navigational effectiveness [1], but future work—possibly observing users directly in a user study or indirectly through log analysis—should indicate what the model's impact, if any, is on navigational effectiveness.

References

[1] E. Meij and M. de Rijke. Thesaurus-based feedback to support mixed search and browsing environments. In *ECDL '07*, 2007.
[2] E. Meij, D. Trieschnigg, M. de Rijke, and W. Kraaij. Parsimonious concept modeling. In *SIGIR '08*, 2008.
[3] E. Meij, W. Weerkamp, K. Balog, and M. de Rijke. Parsimonious relevance models. In *SIGIR '08*, 2008.
[4] D. Trieschnigg, E. Meij, M. de Rijke, and W. Kraaij. Measuring concept relatedness using language models. In *SIGIR '08*, 2008.

SIGIR'08, July 20–24, 2008, Singapore.
ACM 978-1-60558-164-4/08/07.