Beyond Bags of Words: Effectively Modeling Dependence and Features in Information Retrieval

Donald Metzler*
Computer Science Building
140 Governors Drive
University of Massachusetts
Amherst, MA 01003-9264
metzler@cs.umass.edu
http://ciir.cs.umass.edu/∼metzler/metzler-thesis.pdf

Summary. Current state of the art information retrieval models treat documents and queries as bags of words. There have been many attempts to go beyond this simple representation. Unfortunately, few have shown consistent improvements in retrieval effectiveness across a wide range of tasks and data sets. Here, we propose a new statistical model for information retrieval based on Markov random fields. The proposed model goes beyond the bag of words assumption by allowing dependencies between terms to be incorporated into the model. This allows for a variety of textual and non-textual features to be easily combined under the umbrella of a single model. Within this framework, we explore the theoretical issues involved, parameter estimation, feature selection, and query expansion. We give experimental results from a number of information retrieval tasks, such as ad hoc retrieval and web search.

Contributions. The primary contributions of the thesis are:

1. **Robust retrieval model.** We develop a new, formally motivated, statistical retrieval model based on Markov random fields that robustly and effectively handles term dependencies and the combination of arbitrary features.
2. **Better understanding of features for information retrieval.** By modeling dependencies between terms and encoding rich features, such as those based on phrases and term proximity, we are able to better understand how and when such features can improve retrieval effectiveness.
3. **Novel parameter estimation technique.** Our technique exploits the nature of rank-equivalence and works to directly maximize the underlying retrieval metric, which leads to better performance than maximizing the data likelihood or margin. This avoids the problem of metric divergence.
4. **Automatic model learning.** We propose a supervised feature selection algorithm that can be used to automatically learn highly effective models. This eliminates the need for human experts to manually select model features on a per-task basis.
5. **Concept-based query expansion.** Our retrieval model provides an elegant mechanism for expanding queries using multi-term concepts in the context of relevance or pseudo-relevance feedback.
6. **State of the art retrieval effectiveness.** Our model shows consistent and significant improvements in retrieval effectiveness over current state of the art retrieval models on *ad hoc* retrieval and web search tasks.

*Author's current affiliation: Yahoo! Research