Risk Minimization and Language Modeling in Text Retrieval
Dissertation Abstract

ChengXiang Zhai (Advisor: John Lafferty)

Language Technologies Institute
School of Computer Science
Carnegie Mellon University

With the dramatic increase in online information in recent years, text retrieval is becoming increasingly important. Although many different text retrieval approaches have been proposed and studied in the past decades, it is still a significant scientific challenge to develop principled retrieval approaches that also perform well empirically; so far, the theoretically well-motivated models have rarely led to good performance directly. It is also a great challenge in retrieval to develop models that may go beyond the traditional notion of topical relevance and capture more user factors, such as topical redundancy and sub-topic diversity.

This thesis presents a new text retrieval framework based on Bayesian decision theory. The framework unifies several existing retrieval models, including the language modeling approach proposed recently, within one general probabilistic framework, and facilitates the development of new principled approaches to text retrieval with the potential of going beyond the traditional notion of topical relevance. In this framework, queries and documents are modeled using statistical language models (i.e., probabilistic models of text), user preferences are modeled through loss functions, and retrieval is cast as a risk minimization problem.

While traditional retrieval models rely heavily on ad hoc parameter tuning to achieve satisfactory retrieval performance, the use of language models in the risk minimization framework makes it possible to exploit statistical estimation methods to improve retrieval performance and set retrieval parameters automatically. As a special case of the framework, we present a two-stage language model that, according to extensive evaluation, achieves excellent retrieval performance without any ad hoc parameter tuning.

Using language models in retrieval also makes it possible to improve retrieval performance through using improved language models and estimation methods. As another special case of the risk minimization framework, we derive a Kullback-Leibler divergence retrieval model that can exploit feedback documents to improve the estimation of query models. Feedback has so far been dealt with heuristically in the language modeling approach to retrieval.

The KL-divergence model provides a more natural way of performing feedback by treating it as query model updating. We propose two specific query model updating

algorithms based on feedback documents. Evaluation indicates that both algorithms are effective for feedback.

The risk minimization retrieval framework further allows for incorporating user factors beyond the traditional notion of topical relevance. We present language models that can capture redundancy and sub-topics in documents, and study loss functions that can rank documents in terms of both relevance and sub-topic diversity. Evaluation shows that the proposed language models can effectively capture redundancy and can outperform the relevance-based ranking method for the aspect retrieval task.