Improving the Effectiveness of Language Modeling Approaches to Information Retrieval: Bridging the Theory-Effectiveness Gap

Yuanhua Lv
Microsoft Research, Mountain View, CA, USA
yuanhual@microsoft.com

October 7, 2012

Abstract

Improving the effectiveness of general retrieval models has been a long-standing difficult challenge in information retrieval research, yet is also a fundamentally important task, because an improved general retrieval model would benefit every search engine. The language modeling approach to information retrieval has recently attracted much attention. In the language modeling approach, we assume that a query is a sample drawn from a language model: given a query Q and a document D, we compute the likelihood of "generating" query Q with a document language model estimated based on document D. We can then rank documents based on the likelihood of generating the query, i.e., query likelihood. On the one hand, with sound statistical foundation, the language modeling approach makes it easier to set and optimize retrieval parameters, and often outperforms traditional retrieval models. On the other hand, however, after more than one decade of research, the basic language modeling approach to retrieval still remains the same, mainly because the difficulty in accurately modeling the highly empirical notion of relevance within a standard statistical model has led to slow progress in optimizing language modeling approaches; this suggests that the theoretical framework of language models has a clear gap from what is needed to make a retrieval model empirically effective, a general problem we refer to as the "theory-effectiveness gap". We have identified the following theory-effectiveness gaps in current language modeling approaches:

First, one critical common component in any language modeling approach is the document language model. Traditional document language models follow the bag-of-words assumption that assumes term independence and ignores the positions of the query terms in a document. For example, in a query "computer virus", the occurrences of two query terms may be close to each other in one document (likely to mean computer virus) while far apart in another document (not necessarily about computer virus), which makes a huge difference for indicating relevance but is largely underexplored, suggesting the existence of a theory-effectiveness in standard document language models.

Second, accurate estimation of query language models plays a critical role in the language modeling approach to information retrieval. Pseudo-relevance feedback (PRF) has proven very effective for improving query language models. The basic idea of PRF is to assume that

a small number of top-ranked documents in the initial retrieval results are relevant and select from these documents useful terms to improve the query language model. However, existing PRF algorithms simply assume that all terms in a feedback document are equally useful, again ignoring term occurrence positions. They are often non-optimal, as a feedback document may cover multiple incoherent topics and thus contain many useless or even harmful terms. This shows a theory-effectiveness gap in estimating query language models based on PRF.

Third, although pseudo-relevance feedback approaches to the estimation of query language models can help improve the average retrieval precision, many experiments have shown that PRF often hurts many individual queries; the risk of PRF limits its usefulness in real search engines – another theory-effectiveness gap in query language models.

Fourth, the language modeling approach scores a document mainly based on the query likelihood score. A previously unknown deficiency of the query likelihood scoring function is that it is not properly lower-bounded for long documents. As a result of this deficiency, long documents which do match the query term can often be scored unfairly as having a lower relevancy than shorter documents that do not contain the query term at all. For example, for the aforementioned query "computer virus", a long document matching both "computer" and "virus" can easily be ranked lower than a short document matching only "computer". This reveals a clear theory-effectiveness gap between the standard query likelihood scoring function and the optimal way of scoring documents.

Fifth, the justification of using the basic query likelihood score for retrieval requires an unrealistic assumption, which states that the probability that a user who dislikes a document would use a query does not depend on the particular document. In reality, however, this assumption does not hold because a user who dislikes a document would more likely avoid using words in the document when posing a query. This theoretical gap between the basic query likelihood retrieval function and the notion of relevance suggests that the basic query likelihood function is a potentially non-optimal retrieval function.

To bridge the above theory-effectiveness gaps between the theoretical framework of standard language models and the empirical application of information retrieval, in this dissertation, we clearly identified the causes of these gaps, and developed general methodologies to remove the causes from language models without destroying the statistical foundation and any other desirable properties of language models. Our explorations have delivered several more effective and robust general language modeling approaches, which can all be applied immediately to search engines to improve their ranking accuracy. Although this dissertation focuses on language models, most of the proposed methodologies are actually more general, and can also be applied to retrieval models other than language models to bridge their theory-effectiveness gap as well.

This dissertation was completed at the Department of Computer Science at University of Illinois at Urbana-Champaign under the advise of Dr. ChengXiang Zhai. Dr. ChengXiang Zhai, Dr. Jiawei Han, Dr. Evgeniy Gabrilovich, and Dr. Miles Efron served as dissertation committee members. Available online at: https://www.ideals.illinois.edu/handle/2142/34306