

## Abstract

Information retrieval algorithms attempt to match a user's description of their information need with relevant information in a collection of documents or other data. Applications include Web search engines, filtering and recommendation systems, computer-assisted language tutors, and many others. A key challenge of retrieval algorithms is to perform effective matching when many factors, such as the user's true information need, may be highly uncertain and can only be partially observed via a small number of keywords. This dissertation proposes a new research direction for managing this risk by measuring and exploiting the sensitivity of retrieval algorithms to improve their performance. Our contributions include new theoretical models, analytical methods, and retrieval algorithms.

As an application, we focus on a long-studied approach to improving retrieval matching that adds related terms to a query – a process known as query expansion. Query expansion works well on average, but even state-of-the-art methods are still highly unreliable and can greatly hurt results for individual queries. We show how using sensitivity information can significantly improve the reliability of query expansion algorithms without reducing their overall effectiveness, while making few assumptions about the nature of the base expansion algorithm.

Our approach proceeds in two steps. First, treating the base expansion method as a 'black box', we gather information about how the algorithm's output – a set of expansion term weights – changes with perturbations to the initial query and top-ranked documents. This step also results in a set of plausible expansion model candidates. Second, by casting robust query expansion as a convex optimization problem, we obtain a novel risk framework for combining these candidates that can operate selectively to give a much more reliable version of the original baseline expansion algorithm.

Highlights of our results include:

- \* A new algorithmic framework for estimating more reliable and precise query and document models, based on treating queries and document sets as random variables instead of single observations.
- \* The first significant application and analysis of convex optimization methods for query model estimation in information retrieval.
- \* A new family of statistical similarity measures we call perturbation kernels that are efficient to compute and give context-sensitive word clustering.
- \* The introduction of risk-reward analysis to information retrieval, including tradeoff curves, analysis, and risk measures.
- \* A new general form of query difficulty measure that reflects clustering in the collection as well as the relation between a query and the collection.