Finding People, Papers, & Posts: Vertical Search Algorithms & Evaluation
R. Berendsen

Summary

There is a growing diversity of information access applications. While general web search has been dominant in the past few decades, a wide variety of so-called *vertical search* tasks and applications have come to the fore. Vertical search is an often used term for search that targets specific content. Examples include Youtube video search, Facebook graph search, Spotify music recommendation, product search, expertise retrieval, and scientific literature search. In this dissertation, we study vertical search engines for finding people, finding papers, and finding posts.

In a vertical search application, we typically have some background knowledge about the context in which search is taking place. We may know something about the user population, about the tasks they wish to perform, about their information needs, and about the information objects in the collection we make available to them. This knowledge can inform adaptation of retrieval algorithms and evaluation methodology, to provide a better ranking of information objects, or to organize search results more effectively.

For a people search engine, we analyze and automatically detect the types of information needs searchers have. We also provide methods to deal with the high ambiguity of person names, by finding correct referents for name mentions in search results. We find that, to do this sucessfully for people search results, social media profiles need to be treated seperately from other, regular web pages. A special case of people search is expert finding. The flip-side of that task is expert profiling: ranking knowledge areas for an expert of interest. We evaluate expert profiling algorithms by asking experts to judge profiles we generated for them.

Next, we study a search engine for finding scientific papers in a collection where each paper has been labeled with keywords from a thesaurus with research areas, research methodologies, and classification codes. We show that these annotations can be used to automatically construct ground truth for the optimization and evaluation of algorithms for scientific literature search. Such methods have the potential to reduce the need for manual evaluation of the quality of search engines. We show that with a similar method, ground truth for a completely different search task can be generated: finding microblog posts. Microblog posts, e.g., tweets, contain hashtags, which can be interpreted as annotations about the topic of a tweet. The generated ground truth can be used to train machine learning algorithms for microblog search.

This dissertation showcases the need, as well as many opportunities, to leverage background knowledge in vertical search applications. It provides pointers on how this may be done to aid in understanding user information needs, organizing search results, evaluating retrieval algorithms, and generating ground truth.