Federated Search in Heterogeneous Environments

Jaime Arguello
School of Information and Library Science
University of North Carolina
Chapel Hill, NC 27599
USA
jarguello@unc.edu

2011

Abstract

In information retrieval, federated search is the problem of automatically searching across multiple distributed collections or resources. It is typically decomposed into two subsequent steps: deciding which resources to search (*resource selection*) and deciding how to combine results from multiple resources into a single presentation (*results merging*). Federated search occurs in different environments. This dissertation focuses on an environment that has not been deeply investigated in prior work.

The growing heterogeneity of digital media and the broad range of user information needs that occur in today's world have given rise to a multitude of systems that specialize on a specific type of search task. Examples include search for news, images, video, local businesses, items for sale, and even social-media interactions. In the Web search domain, these specialized systems are called *verticals* and one important task for the Web search engine is the prediction and integration of relevant vertical content into the Web search results. This is known as *aggregated web search* and is the main focus on this dissertation.

Providing a single-point of access to all these diverse systems requires federated search solutions that can support *result-type* and *retrieval-algorithm heterogeneity*. This type of heterogeneity violates major assumptions made by state-of-the-art resource selection and results merging methods and motivates the development of new techniques.

While existing resource selection methods derive evidence exclusively from sampled resource content, the approaches proposed in this dissertation draw on machine learning as a means to easily integrate various different types of evidence. These include, for example, evidence derived from (sampled) vertical content, vertical query-traffic, click-through information, and properties of the query string. In order to operate in a heterogeneous environment, we focus on methods that can learn a vertical-specific relationship between features and relevance. We also present methods that reduce the need for human-produced training data. In particular, we focus on the situation where we have vertical-relevance judgments for some verticals and want to learn a predictive model for a vertical associated with no training data.

Existing results merging methods formulate the task as score normalization. In a more heterogeneous environment, however, combining results into a single presentation requires satisfying a number of layout constraints. The dissertation proposes a novel formulation of the task: *block ranking*. During block-ranking, the objective is to rank sequences of results that must appear grouped together (vertically or horizontally) in the final presentation. Based on this formulation, the dissertation proposes and empirically validates a cost-effective methodology for evaluating aggregated web search results. Finally, it proposes the use of machine learning methods for the task of block-ranking.

This dissertation was completed at School of Computer Science at Carnegie Mellon University under the advise of Dr. Jamie Callan (dissertation committee chair). Dr. Jaime Carbonell, Dr. Yiming Yang, and Dr. Fernando Diaz served as dissertation committee members. For the full dissertation, visit: `http://ils.unc.edu/~jarguell`