

Abstract

Conventional search engines like Google provide access to Web information that can be acquired easily by crawling hyperlinks. However, a large amount of information cannot be copied arbitrarily by conventional search engines. This type of hidden information, which is very valuable, can only be accessed via an alternative search model other than the centralized retrieval model used by conventional search engines.

Federated search provides access to the hidden information by providing a single interface that connects to multiple source-specific search engines. There are three main research problems in federated search. First, information about the contents of each individual information source must be acquired (*resource representation*). Second, given a query, a set of sources must be selected to do the search (*resource selection*). Third, the results retrieved from selected sources may be merged into a single list before it is presented to the end user (*results merging*).

This dissertation addresses these main research problems within federated search. New algorithms are proposed for effectively and efficiently estimating information source sizes, estimating distributions of relevant documents across information sources for a given query, and merging document rankings returned by selected sources. Furthermore, a unified utility maximization framework is proposed to combine the range of individual solutions together to construct effective systems for different federated search applications. The framework can incorporate information such as search engine retrieval effectiveness, which is an important issue for real world federated search applications. Empirical studies in a wide range of research environments and a real world prototype system under different operating conditions demonstrate the effectiveness of the new algorithms.

This new research, supported by a more theoretical foundation, better empirical results, and more realistic simulation of real world applications, substantially improves the state-of-the-art of federated search. It serves as a bridge for moving federated search from an interesting research topic to a practical tool for real world applications.