From Federated to Aggregated Search

Fernando Diaz
Yahoo! Labs, diazf@yahoo-inc.com

Mounia Lalmas
University of Glasgow, mounia@acm.org

Milad Shokouhi
Microsoft Research, milads@microsoft.com

Abstract:

Federated search refers to the brokered retrieval of content from a set of auxiliary retrieval systems instead of from a single, centralized retrieval system. Federated search tasks occur in, for example, digital libraries (where documents from several retrieval systems must be seamlessly merged) or peer-to-peer information retrieval (where documents distributed across a network of local indexes must be retrieved).

In the context of web search, aggregated search refers to the integration of non-web content (e.g. images, videos, news articles, maps, tweets) into a web search result page. This is in contrast with classic web search where users are presented with a ranked list consisting exclusively of general web documents. As in other federated search situations, the non-web content is often retrieved from auxiliary retrieval systems (e.g. image or video databases, news indexes).

Although aggregated search can be seen as an instance of federated search, several aspects make aggregated search a unique and compelling research topic. These include large sources of evidence (e.g. click logs) for deciding what non-web items to return, constrained interfaces (e.g. mobile screens), and a very heterogeneous set of available auxiliary resources (e.g. images, videos, maps, news articles). Each of these aspects introduces problems and opportunities not addressed in the federated search literature.

Aggregated search is an important future research direction for information retrieval. All major search engines now provide aggregated search results. As the number of available auxiliary resources grows, deciding how to effectively surface content from each will become increasingly important.

The goal of this tutorial is to provide an overview of federated search and aggregated search techniques for an intermediate information retrieval researcher. At the same time, the content will be valuable for practitioners in industry. We will take the audience through the most influential work in these areas and describe how they relate to real world aggregated search systems. We will also list some of the new challenges confronted in aggregated search and discuss directions for future work.

ACM Categories & Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Algorithms, Design, Experimentation, Measurement

Keywords

Distributed information retrieval, Federated search, Aggregated search, Universal search, Vertical search, Metasearch

Bios:

Fernando Diaz is a research scientist at Yahoo! Labs. His primary research interests concern formal models of information retrieval, and more recently aggregated search. His research experience also includes distributed information retrieval approaches to web search, interactive and faceted retrieval, mining of temporal patterns from news and query logs, cross-lingual information retrieval, graph-based retrieval methods, and synthesizing information from multiple corpora. He received his PhD from the University of Massachusetts Amherst in 2008.

Mounia Lalmas holds a Microsoft Research/Royal Academy of Engineering Research Chair in Information Retrieval at the Department of Computing Science, University of Glasgow. Her research focuses on the development and evaluation of intelligent access to interactive heterogeneous and complex information repositories. From 2002 until 2007, she co-led the Evaluation Initiative for XML Retrieval (INEX). She is now working on technologies for aggregated search and bridging the digital divide. She is also looking at the use of quantum theory to model interactive information retrieval.

Milad Shokouhi is an applied researcher working for Bing at Microsoft Research Cambridge. Before joining Microsoft in 2007, he did his PhD on federated search at the Royal Melbourne Institute of Technology (RMIT) University. His research interests are federated search, query expansion, user studies and web search evaluation.

SIGIR'10, July 19–23, 2010, Geneva, Switzerland.
ACM 978-1-60558-896-4/10/07.