

Semantic Search

by Darrell W. Gunter

Search 2009

Over the last four decades steady improvements have been made in search engines, including those that use Boolean search interfaces as well as those that rely on other ways of matching query terms to words in the documents searched. In this paper we consider the further gains that can be achieved with a technique called *semantic search*.

Wikipedia defines *semantics* as the study of meaning [1]. The word *semantics* itself denotes a range of ideas, from the popular to the highly technical. It is often used in ordinary language to denote a problem of understanding that comes down to word selection or connotation. Semantic search is a process used to improve online searching by using data from semantic networks [2] to disambiguate queries and web text in order to generate more relevant results.

For example, Collexis [3] utilizes semantic technology to create their proprietary *Fingerprints*. A Fingerprint represents a visualization of the aggregation of the key concepts of the aggregated dataset – presented in a bar chart of relevant weighted concepts. For example, if a researcher has found an article about diabetes in the *New York Times* and wants to know if there is any research in PubMed [4] that matches this article, she can paste the article into the Collexis engine, and it will create a Fingerprint of the article. Further it will return a result set of articles from the PubMed database that can match any portion of the article. In addition to creating Fingerprints and providing matching articles based on a comparison of the relevant concepts, this semantic technology also allows the researcher to profile experts, determine the research trend of key concepts and conduct

hypothesis generation. Further, she is able to build on her search by adding or deleting concepts based on the direction of her research.

Semantic search as described above has many benefits over traditional search. It allows searchers the following capabilities:

- To refine their search by adjusting the settings of the Fingerprint
- To combine or remove a concept(s)
- To create dashboards based on the key relevant concepts

Imagine a search in Google about juvenile diabetes. The result set will contain thousands of items. In order to know the content of all retrieved resources the researcher would have to read them. With the Collexis technology the result set is presented in an aggregated Fingerprint that consists of the weighted bar chart of key relevant concepts. The researcher is able to dive into any concept to determine the articles that are associated with the concept. In a nutshell, the semantic technology allows search based on concepts in context.

These features allow researchers to be more proficient and efficient in their searches. You may know that Microsoft has launched “Bing,” its first semantic entry in the global search space. Bing’s technology came from Microsoft’s acquisition of Powerset last year. Bing has begun to get very good reviews, and I would suspect that Google will not be far behind with a semantic search engine.

Let’s look at a couple of case studies that demonstrate the adaptability and power of this approach. The first case study is from Johns Hopkins University, where they faced a fundamental problem. With hundreds of researchers scattered throughout their institution it was very difficult for the research community to determine the best person to collaborate on a very specific research project. To help with this problem Johns Hopkins

Darrell W. Gunter is EVP/CMO at Collexis Holdings, Inc. He can be reached by email at gunter@collexis.com

Leadership built a centrally located coffee shop – hoping that the research community would get to know their fellow researchers while getting a cup of joe. While everyone enjoyed a great cup of coffee or latte, folks were not making the necessary connections for their research. The university turned to Collexis to help with the expertise identification problem. Collexis was able to aggregate the research profiles of the entire research community and create an institutional dashboard. This institutional dashboard allowed them to immediately find the expert on any biomedical topic or category by searching based on key concepts. Johns Hopkins was so excited about the institutional dashboard that they opened it up to the world via the Internet.

Now its research community can easily identify potential collaborators in a matter of seconds. Its pool of collaborators is also visible to everyone. The coffee shop is still very active, but the new collaborators meet there to discuss research ideas over a cup of joe. It is important to note that this case study was replicated at many top research institutions such as the NIH, the University of Michigan, and the University of Miami to name a few.

The second case study that I would like to discuss is how the American Association of Cancer Research utilizes the Collexis Reviewer Finder to enhance the editor's capability to determine the best pool of peer reviewers. The demands on editors' time are quite significant, and one of their most important duties is to manage the peer review process for their respective journals. Managing the hundreds, if not thousands, of submitted manuscripts is a daunting task. The editors must understand the specifics of the manuscript and then determine who within their peer review pool is best qualified to review it. Today this task is usually managed in a very ad hoc manner. Editors must use a plethora of tools to achieve their objective of getting the manuscripts peer reviewed in a timely and professional manner. To help with this productivity issue, Collexis has developed and launched the Collexis Reviewer Finder application.

Utilizing the Collexis proprietary Fingerprinting technology, the Reviewer Finder applications allow the editor to first Fingerprint the submitted manuscript and then compare it against the 1.8 million research profiles in the BiomedExperts.com database. The Reviewer Finder instantly provides the editor with a list of potential peer reviewers based on the specific

research profile and how the profile dovetails with the submitted manuscript. Further, the Reviewer Finder will also detail if anyone in the peer-reviewer pool has a conflict of interest with the author of the submitted manuscript. A conflict of interest is noted if the potential reviewer has either co-authored a paper/grant or has worked in the same location/institution with the submitting author. The Collexis Reviewer Finder provides editors with an invaluable tool that saves them considerable time and improves their productivity greatly.

Semantic search makes possible the recreation of the two key features in these case studies – aggregating and visualizing a large dataset and conducting a conceptual search. Standard search tools do not support these tasks in a very precise or timely way. While they will continue to be used, it is time for new technology to step forward that can increase the researcher productivity. I am confident that this new technology is semantic search. By this time next year we will see a flood of exciting applications based on semantic technology. For more information about the Semantic Web visit the Project10x website [5] and download the executive summary for *Semantic Wave Report: Industry Roadmap to Web 3.0 and Multibillion Dollar Market Opportunities* [6]. ■

Resources Mentioned in the Article

- [1] Semantics. *Wikipedia.org*. Retrieved August 20, 2009, from <http://en.wikipedia.org/wiki/Semantics>.
- [2] Semantic networks. *Wikipedia.org*. Retrieved August 20, 2009, from http://en.wikipedia.org/wiki/Semantic_networks.
- [3] Collexis [website]: www.collexis.org.
- [4] PubMed: www.ncbi.nlm.nih.gov/pubmed.
- [5] Project 10x: www.project10x.com.
- [6] Davis, M. (n.d.). *Semantic wave report: Industry roadmap to Web 3.0 and multibillion dollar market opportunities: Executive summary*. Washington, DC: Project10x. Retrieved August 21, 2009, from www.project10x.com.