

Surfing the Wave or Taking a Road Less Traveled: The 2008 Infonortics Search Engine Meeting

by Paul Thompson

The 2008 Infonortics Search Engine Meeting was held in Boston on April 28-29. As in the past, this meeting included talks and vendor booths from commercial organizations and universities. Most of the PowerPoint presentations are available at the conference website: www.infonortics.com/searchengines/sh08/08pro-post.html.

The conference opened with Charles Clarke of the University of Waterloo, Canada, describing XML retrieval research representative of what is presented at conferences such as INEX (Initiative for the Evaluation of XML). While this paper was voted the best paper of the conference, it was one of the least typical of the search engine meeting. More typical papers described commercial search products. The second paper, presented by Stephen Arnold, described new semantic web technologies being developed by Google. He claims that Google is cornering this market and that would-be competitors, rather than attempting to compete directly, should instead surf the wave made by Google.

Enterprise Search and Business Intelligence

Most of the presentations were on enterprise search and business intelligence. Within this category several subcategories were present, though often a company's presentation fit in more than one category. Several presentations, notably the one given by Abe Lederman of Deep Web

Paul Thompson has had an extensive career researching and implementing search engine technology. He is presently engaged in information extraction and retrieval research at Dartmouth College. He can be reached at Paul.Thompson@Dartmouth.edu

Technologies, addressed the problem of accessing the deep or hidden web. ISYS' presentation described the iterative nature of search and the need to understand different types of users and information needs. Edwin Cooper of InQira made a similar point, of which more will be said below.

Several presenters described technologies that allowed structured and unstructured data to be searched together. Sid Probst of Attivio focused on this point. Spencer Shearer of Exalead added to the discussion of structured and unstructured data the concept of hybrid search, utilizing mashup technology. He claims that using mashups would fill the large search gaps missed by both structured and unstructured search.

Other business intelligence presentations described different variations of semantic search. Pascal Coupet of Temis emphasized the importance of accurate information extraction and described technology that enabled users to correct errors in automatic extraction. Roger Bradford of Agilex Technologies discussed sentiment detection. Kelly Stirman of Mark Logic also described sentiment detection, but based on support-vector-machine technology. Sam Chapman of the University of Sheffield described a commercial spin-off, K-now, based on his university's extensive research and development in the area of natural language understanding. This technology combines keyword and semantic approaches to search. George Chitouras of Business Objects also discussed combining information retrieval and natural language technologies. Jeff Fried of Microsoft spoke on the confluence of search and business intelligence. Brad Allen of Siderean Software presented a different approach to semantic search, based more on Web 2.0 social computing than on natural language technologies.

Several of the business intelligence presentations also considered the importance of visualization. Marcelline Saunders of Groxis gave an excellent overview of the use of visualization technologies to support information retrieval. Richard Brath of Oculus gave a presentation on visualization and sense-making. He emphasizes the need for a more integrated and holistic approach to support analysts, including a mixed-initiative framework that supports the whole workflow. This talk received the second best paper award.

Miscellaneous Presentations

Several interesting presentations were made on topics other than enterprise search or business intelligence. Jason Baron of the National Archives described the e-discovery problem in the legal domain, including the related legal track at the TREC conference. Peter Jackson of Thomson-Reuters described a recommender system used with legal searches. Steven Forth and Amelia Newbury of the Monitor Group spoke about the role of search in corporate training. Chris Cleveland of Dieselpoint described Open Pipeline, an open source environment for developing search applications.

Web Search vs. Enterprise Search

Edwin Cooper's presentation, "Two Roads Diverged in a Google World," was of particular interest. He notes that enterprise search technology is often seen as a lagging derivative of web search technology. He argues, however, that web search and enterprise search are fundamentally different. For example, web search must accommodate a wide variety of users with many different types of needs, whereas enterprise search is much more constrained. As both web and enterprise search technologies improve, adapting to these very different requirements will inevitably result in web and enterprise search technologies increasingly diverging.

Summary

Arnold recommends that search engine vendors ride the Google wave. But if the presentations at this meeting are any guide, vendors do not appear to be following this course. Cooper's analysis of the difference between web and enterprise search may help explain why the field of search remains so diverse and why the web-scale approaches taken by Google leave much room for other vendors in enterprise search and business intelligence. ■