

Introduction to *Search 2009*

by Marjorie M.K. Hlava, Guest Editor, and Jay Ven Eman

Search 2009

Search is like the blind men and the elephant. The story goes that four blind men approached an elephant. By touching the hide, trunk, tail and legs each thought of the elephant as an entirely different animal. From the hide we can conclude that search is a contextual process where searchers must aimlessly feel around until they either accidentally find what they want or give up. From the trunk we can conclude that search should be like a fire hose where we flood the searchers with content, overwhelming them. The tail? From this perspective search is seen as the infamous “long tail” wherein searchers are fed minuscule tidbits of esoterica, leaving them curious, but famished. And those feeling the legs would conclude that we should stomp the content into submission, pulverizing until it is unrecognizable.

The make-up of our blind crew is further complicated because it represents software developers, entrepreneurs, content producers and searchers – different perspectives and different needs. And each has different needs at different points in time. If we focus on just the searchers, their needs change over time. Search behavior is not uniform across a given population of searchers and any individual’s search behavior will change with his needs. Information requirements at the beginning of a project are likely to change through the course of a project. The serendipity created by browsing the stacks – or dynamically generated clusters produced from a search query – will likely be appreciated at the onset of a knowledge intensive project.

Marjorie M.K. Hlava, guest editor of this special section of the *Bulletin of the American Society for Information Science and Technology*, is president and chairman of Access Innovations, Inc., and a former president of ASIS&T. She can be reached at mhlava@accessinn.com.

Jay Ven Eman, co-author of this introduction, is CEO of Access Innovations, Inc. He can be reached at j_ven_eman@accessinn.com.

Once the project parameters are set and the project approved, precision and recall in retrieval are paramount. For a typical project timeline stage would be about 95% of its duration.

To compound the problem, Google has made search known. Prior to Google’s ascendance, search was not really known outside the information industry, where it was the purview of information professionals serving professional knowledge workers. Now search is everywhere. Google has also biased the landscape of expectations with little differentiation given to the World Wide Web versus large, complex, internal information environments – the world of filing cabinets, billions of C drives and an equal number of desks stuffed with paper documents, and of electronic content in a near infinite number of formats, spread across thousands of applications.

The elephant parable illustrates the complexity that is search and helps explain why search is so frustrating and doesn’t work most of the time. It helps explain why the typical, large organization has, on average, five separate search applications. It also highlights why there are so many different approaches to solving search.

What are the basic approaches to search to ensure that end users will be able to find their data easily, quickly and accurately? The kinds of search algorithms used to build and implement search software systems vary widely. There are Boolean engines, Bayesian engines, inference, vector, ranking, natural language processing (NLP) and its parts (semantic, syntactic, phraseological, morphological, grammatical, common sense), co-occurrence, clustering, sequel rules, neural nets, latent semantic and others, not to mention various combinations thereof.

Most search applications now use more than one approach in an attempt to overcome various weaknesses inherent in each approach as well as trying to complement strengths. Whether this will result in fuller, richer, more

satisfying experiences for searchers or add to their frustration and to the complexity and cost of implementation and maintenance is yet to be determined for many of the offerings in the marketplace today. There is not much research on whether combined solutions are necessarily any better than single design approaches. Does combining latent semantic indexing with NLP compound the weaknesses of each, or is the whole better than its parts?

In modern search technologies much is done to surround search to bring a good answer to the requestor. This issue considers search on the practical side, using case studies to illustrate varying points of view. Some of the matters covered include dealing with unstructured information, meaning text that is not field-formatted into relational databases, but rather has been left as a full-text document in formats such as PDF or the various formats included in Microsoft Office. Other approaches suggest semantic and structural enrichment of the content as a critical part of the search solution.

Measuring search results by relevance, precision and recall (or hit, miss and noise statistics) has been discussed in detail in the *Journal of the American Society for Information Science and Technology (JASIST)* from a theoretical perspective. Visualization of the results is also an important topic, but in this issue we focus on what search software developers are doing to produce better results metrics. Pragmatic case studies look at the options and the challenges of implementation at an organization's site. The authors were chosen to represent the breadth of actual approaches from Boolean to latent semantic and the wide territory in between.

Search challenges have been with us a long time. At the turn of the 1900s, Charles Ami Cutter and Melville Dewey took very different approaches to the way in which items should be stored and retrieved. Cutter [1] leaned toward multiple classifications and heavy cross-referencing to "make the hoard an army," while Dewey preferred single placement and detailed classifications with see also references [2].

In the 1930s, GE Research under innovators like Homer Hall was able to use early computer logic to organize information resources. [3]

The Cold War brought us the massive research of COSATI (Council on Scientific and Technical Information, 1964) [4, 5], leading to the Recon and later Dialog systems, which dominated our world for two decades. [6, 7, 8]

Many of their recommendations are still waiting to be implemented. At the same time the SDC Orbit system and the ELHILL work at the National Library of Medicine took another approach. [9] These Boolean systems were the vanguard of the 1990s.

Recently we have seen a profusion of search software from university researchers: from Autonomy, growing out of research at the University of Cambridge, to the Google cluster of algorithms and computer hardware power, born from the Stanford University research. The discussion continues as to whether it is better a) to add value to data with controlled vocabularies and metadata or b) let the software, properly tuned, do it all with no need for human intervention. The truth, as those blind men found, is the whole is usually greater than the sum of its parts – there will not be "The One Solution."

We once worried about reaching the "elusive end user." Now, with the advent of Internet protocols and reaching near ubiquitous accessibility, everyone has begun to search. The challenges faced in the last century are still with us. The papers in this issue exemplify some best-of-breed directions from practical perspectives. All of the authors are involved in commercially viable systems. Their pragmatic approaches are, by necessity, steeped in the results of search research. Meetings like the ASIS&T Information Architecture Summita address the frontiers of user interface design. The underlying architecture is discussed in this issue.

The article by John Holt takes us through the evolution of search to its current level. Similarly, Rich Turner takes us on an exploration of next generation search platforms. The article by Daniel Tunkelang outlines the options for relevance and ways to measure the results. Rafsky takes us one step further demonstrating through a case study the challenges encountered in building a search service for business and financial news including real-time delivery and the need to measure the impact for each returned search and to bridge the gap from system-based to user-based ranking.

The differentiation between the search presentation layer, the user interface and the search software underneath is not always apparent. The article by Ronald Millett makes this distinction clear and provides examples of the data on two different search software applications with the same overlying presentation layer. He explores the relationships between taxonomy,

metadata, fields and facets, and the ways in which taxonomy can work with search. In the end, there is always an inverted file to associate the query with the data. Most systems separate the display files from the search files to speed the response time for the user. The article by Darrell Gunter shows the levels of options available, including personalization through search.

Here's an article-by-article preview of the contents of this issue:

John D. Holt and David J. Miller, "An Evolution of Search"

A new stage in the evolution of search has arrived with the advent of entity-based search. The linking or clustering of documents or records into sets of references that describe an entity can be used for much more than just reporting on an entity. The entity can become the object of the search. Entity search builds upon Boolean and relevance ranking techniques, providing improvements in both precision and recall. Precision is improved because the entities returned are consistent with the attribute values supplied in the search. Recall is improved because the combination of entity values specified in the search expression need not appear in any particular underlying reference document or record.

Rich Turner, "Next Generation Search Platforms: How Vendors Are Searching Unstructured Content"

Search software companies are looking to develop the next generation technologies that will drive past keyword search by understanding the concepts within unstructured information. Three technologies are receiving wide attention: natural language processing, Bayesian inference and latent semantic indexing and analysis. This article examines all three technologies, exploring what they do and the pros and cons behind each technique.

Daniel Tunkelang, "Reconsidering Relevance and Embracing Interaction"

The query does not provide search engines with enough information to reliably determine how relevant a document is to users' information needs. Human-computer information retrieval (HCIR) tools provide users with opportunities to clarify and elaborate their intent. If the engine isn't sure

what users want, it can ask them. The author considers the three goals of transparency, control and guidance.

Lawrence C. Rafsky, "Searching Real-Time Financial News: PAR for the Course"

Many challenges are encountered in building a search service that includes real-time delivery for business and financial news. PAR—Presumed Action Response is built to measure the impact for each returned search and bridge the gap from system-based to user-based ranking.

Ronald Millett, "Integration of Taxonomy and Keyword Searches: A Comparison of Two Implementations"

Two search implementations, one Bayesian and one Boolean, which combine taxonomy and keyword search using the same database and thesaurus are compared. The advantages of including taxonomy-based, subject descriptor searches and keyword searches in the same system were seen in both test systems. More relevant results were achieved through tapping multiple dimensions of search. The Search Harmony system combines several search methodologies in one, unified system. These methodologies include navigational trees, subject descriptors, auto-completion, search within results and relevance-based full-text searches including fuzzy search features such as stemming. Combining all of these features in a single interface improves search results.

Darrell W. Gunter, "Semantic Search"

The author offers two case studies to support the exploration of the differences between traditional Boolean search and semantic search. Semantic search is a process used to improve online searching by using data from semantic networks to disambiguate queries and web text in order to generate more relevant results. Using semantic technology creates proprietary aggregated visualizations of the key concepts of the aggregated dataset – presented in a bar chart of relevant weighted concepts. The contention is that this allows researchers to be more proficient and efficient in their search. ■

RESOURCES on next page

Resources Mentioned in the Article

- [1] Cutter, C.A. (1876). *Rules for a dictionary catalog: Public libraries in the United States of America: Their history condition and management: Special report. Part II.* Washington, D.C.: Government Printing Office. Retrieved September 2, 2009, from http://books.google.com/books?id=rj-f4-Ps-AkC&dq=Charles+Ammi+Cutter&printsec=frontcover&source=an&hl=en&ei=1sadSvyyBoiYMZn8lJEC&sa=X&oi=book_result&ct=result&resnum=6#v=onepage&q=&f=false.
- [2] Dewey, M. (1899). *Decimal classification system and relativ index for libraries: Clippings, notes, etc.* Boston: Library Bureau. Retrieved September 2, 2009, from http://books.google.com/books?id=wCMbAAAAMAAJ&dq=Melvil+Dewey&printsec=frontcover&source=bl&ots=Hthvuya7ga&sig=d0Jx-5sUXm9hCdlZKCRixNPr3qs&hl=en&ei=OcidSqyqNleCMsTYyYkC&sa=X&oi=book_result&ct=result&resnum=14#v=onepage&q=&f=false.
- [3] Neils Bohr Library and Archives, & The Center for the History of Physics. *Finding Aid to the Records of the Massachusetts Institute of Technology. Office of the President, 1930-1959.* College Park, MD: American Institute of Physics. Center for the History of Physics. Retrieved September 2, 2009, from www.aip.org/history/ead/19990046.html
- [4] American Documentation Institute. Committee on Organization of Information. (1965). Report for 1963-1964 of the Committee on Organization of Information to American Documentation Institute. *American Documentation* 16(3), 235-241. Abstract retrieved September 2, 2009, from www3.interscience.wiley.com/journal/114214118/abstract
- [5] President's Science Advisory Committee. (January 10, 1963). *Science, government, and information: The responsibilities of the technical community and the government in the transfer of information.* Washington, DC: The White House [Also known as the *COSATI Report* or the *Weinberg Report* (after Alvin Weinberg, the Committee chairman).]
- [6] Dialog. (n.d.) *The history of Dialog – Transcript.* Morrisville, NC: Dialog. Retrieved September 2, 2009, from www.dialog.com/about/history/transcript.shtml
- [7] Bourne, C. P., & Hahn, T.B. (2003.) *A history of online information services, 1963 – 1976.* Cambridge, MA: MIT Press.
- [8] Newman, Donald. (January 1986). Dialogue on Dialog: Interview with Roger Summit (EJ328434), *Wilson Library Bulletin*, 60(5), 21-25. Abstract retrieved September 2, 2009, from www.eric.ed.gov/ERICWebPortal/custom/portlets/recordDetails/detailmini.jsp?_nfpb=true&_ERICExtSearch_SearchValue_0=EJ328434&ERICExtSearch_SearchType_0=no&accno=EJ328434
- [9] Remembering ELHILL. (July-August 1999). *NLM Technical Bulletin*, 309. Retrieved September 2, 2009, from www.nlm.nih.gov/pubs/techbull/ja99/ja99_remember.html.