



Beyond Search

News and Information from ArnoldIT.com
about search and content processing...

Beyond Search

WEDNESDAY, FEBRUARY 18, 2009

Exclusive Interview with Kathleen Dahlgren, Cognition Technologies

February 18, 2009

[Cognition Technologies](#)' Kathleen Dahlgren spoke with Harry Collier about her firm's search and content processing system. Cognition's core technology, Cognition's Semantic NLPTM, is the outgrowth of ideas and development work which began over 23 years ago at IBM where Cognition's founder and CTO, Kathleen Dahlgren, Ph.D., led a research team to create the first prototype of a "natural language understanding system." In 1990, Dr. Dahlgren left IBM and formed a new company called Intelligent Text Processing (ITP). ITP applied for and won an innovative research grant with the Small Business Administration. This funding enabled the company to develop a commercial prototype of what would become Cognition's Semantic NLP. That work won a Small Business Innovation Research (SBIR) award for excellence in 1995. In 1998, ITP was awarded a patent on a component of the technology.

Dr. Dahlgren is one of the featured speakers at the Boston Search Engine Meeting. This conference is the world's leading venue for substantive discussions about search, content processing, and semantic technology. Attendees have an opportunity to hear talks by recognized leaders in information retrieval and then speak with these individuals, ask questions, and engage in conversations with other attendees. You can get more information about the Boston Search Engine Meeting [here](#).

The full text of Mr. Collier's interview with Dr. Dahlgren, conducted on February 13, 2009, appears below:

Will you describe briefly your company and its search / content processing technology?

CognitionSearch uses linguistic science to analyze language and provide meaning-based search. Cognition has built the largest semantic map of English with morphology (word stems such as catch-caught, baby-babies, communication, intercommunication), word senses (strike meaning hit, strike a state of baseball, etc.), synonymy ("strike" meaning hit, "beat" meaning hit, etc.), hyponymy ("vehicle"- "motor vehicle"- "car"- "Ford"), meaning contexts ("strike" means game state in the context of "baseball") and phrases ("bok-choy"). . The semantic map enables CognitionSearch to unravel the meaning of text and queries, with the result that search performs with over 90% precision and 90% recall.

What are the three major challenges you see in search / content processing in 2009?

That's a good question. The three challenges in my opinion are:

1. Too much irrelevant material retrieved – poor precision
2. Too much relevant material missed – poor recall
3. Getting users to adopt new ways of searching that are available with advanced search technologies. NLP semantic search offers users the to state longer queries in plain English and get results, but they are currently used to keywords, so there will be an adaptation required of them to take advantage of the new advanced technology.

With search / content processing decades old, what have been the principal barriers to resolving these challenges in the past?

Poor precision and poor recall are due to the use of pattern-matching and statistical search software. As long as meaning is not recovered, the current search engines will produce mostly irrelevant material. Statistics on popularity boost many of the relevant results to the top, but as a measure across all retrievals, precision is under 30%. Poor recall means that sometimes there are no relevant hits, even though there may be many hits. This is because the alternative ways of expressing the user's intended meaning in the query are not understood by the search engine. If they add synonyms without first determining meaning, recall can improve, but at the expense of extremely poor precision. This is because all the synonyms of an ambiguous word in all of its meanings, are used as search terms. Most of these are off target. While the ambiguous words in a language are relatively few, they are among the most frequent words. For example, the seventeen thousand most frequent words of English tend to be ambiguous.

What is your approach to problem solving in search and content processing?

Cognition focuses on improving search by improving the underlying software and making it mimic human linguistic reasoning in many respects. CognitionSearch first determines the meanings of words in context and then searches on the particular meanings of search terms, their synonyms (also disambiguated) and hyponyms (more specific word meanings in a concept hierarchy or ontology). For example, given a search for "mental disease in kids" CognitionSearch first determines that "mental disease" is a phrase, and synonymous with an ontological node, and that "kids" has stem "kid", and that it means "human child" not a type of "goat". It then finds document with sentences having "mental-disease" or "OCD" or "obsessive compulsive disorder" or "schizophrenia", etc. and "kid" (meaning human child) or "child" (meaning human child) or "young person" or "toddler", etc.

Multi core processors provide significant performance boosts. But search / content processing often faces bottlenecks and latency in indexing and query processing. What's your view on the performance of your system or systems with which you are familiar?

Natural language processing systems have been notoriously challenged by scalability. Recent massive upgrades in computer power have now made NLP a possibility in web search. CognitionSearch has sub-second response time and is fully distributed to as many processors as desired for both indexing and search. Distribution is one solution to scalability. Another CognitionSearch implements is to compile all reasoning into the index, so that any delays caused by reasoning are not experienced by the end user.

Google has disrupted certain enterprise search markets with its appliance solution. The Google brand creates the idea in the minds of some procurement teams and purchasing agents that Google is the only or preferred search solution. What can a vendor do to adapt to this Google effect? Is Google a significant player in enterprise search, or is Google a minor player?

Google's search appliance highlights the weakness of popularity-based searching. On the web, with Google's vast history of searches, popularity is effective in positioning the more desired sites at the top the relevance rank. Inside the enterprise, popularity is ineffective and Google performs as a plain pattern-matcher. Competitive vendors need to explain this to clients, and even show them with head-to-head comparisons of search with Google and search with their software on the same data. Google brand allegiance is a barrier to sales in enterprise search.

Information governance is gaining importance. Search / content processing is becoming part of eDiscovery or internal audit procedures. What's your view of the the role of search / content processing technology in these specialized sectors?

Intelligent search in eDiscovery can dig up the "smoking gun" of violations within an organization. For example, in the recent mortgage crisis, buyers were lent money without proper proof of income. Terms for this were "stated income only", "liar loan", "no-doc loan", "low-documentation loan". In eDiscovery, intelligent search such as CognitionSearch

would find all mentions of that concept, regardless of the way it was expressed in documents and email. Full exhaustiveness in search empowers lawyers analyzing discovery documents to find absolutely everything that is relevant or responsive. Likewise, intelligent search empowers corporate oversight personnel, and corporate staff in general, to find the desired information without being inundated with irrelevant hits (retrievals). Dedicated systems for eDiscovery and corporate search need only house the indices, not the original documents. It should be possible to host a company-wide secure Web site for internal search at low cost.

As you look forward, what are some new features / issues that you think will become more important in 2009? Where do you see a major break-through over the next 36 months?

Semantics and the semantic web have attracted a great deal of interest lately. One type of semantic search involves tagging of documents and Web sites, and relating them to each other in a hierarchy expressed in the tags. This type of semantic search enables taggers to perfectly control reasoning with respect to the various documents or sites, but is labor-intensive. Another type of semantic search is runs on free text, is fully automatic, and uses semantically-based software to automatically characterize the meaning of documents and sites, as with CognitionSearch.

Mobile search is emerging as an important branch of search / content processing. Mobile search, however, imposes some limitations on presentation and query submission. What are your views of mobile search's impact on more traditional enterprise search / content processing?

Mobile search heightens the need for improved precision, because the devices don't have space to display millions of results, most of which are irrelevant.

Where can I find more information about your products, services, and research?

<http://www.cognition.com>

Harry Collier, [Infonortics, Ltd.](#), February 18, 2009
Written by Stephen E. Arnold ·
