Science | DOI:10.1145/1629175.1629183 Neil Savage

New Search Challenges and Opportunities

If search engines can extract more meaning from text and better understand what people are looking for, the Web's resources could be accessed more effectively.

The Web is a huge, dynamic landscape of information, and navigating through it not an easy task. There are billions of Web pages, and the type of content is expanding dramatically, with blogs and Twitter feeds, maps and videos, photos and podcasts. People, typing on a computer in their cubicle or using their smartphone on a street corner, are trying to sift through this growing morass of data, looking for everything from car repair advice to a nearby Thai restaurant that's not too expensive. For search engines, this enormous variety of data and formats is providing both new challenges and new opportunities.

"The ability to produce information and store information has far outstripped human cognitive capacity, which is basically fixed," says Oren Etzioni, a professor of computer science and engineering at the University of Washington. "The haystack keeps getting bigger. Obviously we need better and better tools to find the proverbial needles."

Today's search engines do a fine job of cataloging text, counting links, and delivering lists of pages relevant to a user's search topic. But in the coming decade, Etzioni believes, search will move beyond keyword queries and automate the time-consuming task of sifting through those documents. With a better understanding both of what documents mean and what searchers are looking for, he predicts, some tasks could be reduced from hours to minutes.

Etzioni is attempting to get more information out of text using a technique called open information extraction, which is built on a long-used technology that examines natural language text and tries to derive data about the relationships between words. An algorithm looks for triples, which follow the structure of entity-relationship-entity, such as "Beijing is the capital of China" or "Franz Kafka was born in Prague." The system is open because it derives the relations from the structure of the language rather than relying on hand-labeled examples of relationships, which would not be scalable to the Web as a whole.

Like many other computer scientists, the University of Washington's Oren Etzioni is developing new tools for searching the Web's growing morass of text, images, and other content.

Oren Etzioni's approach examines natural language text and tries to derive data about the relationships between words.

Etzioni developed a program called TextRunner that uses a general model of language to assign labels to words in a sentence, then to calculate the beginning and end of strings of words that contain the entity-relationship-entity structure. It extracts those triples so they can be indexed and searched. A searcher who asks "Where was Kafka born?" should quickly receive a precise answer, not just a list of pages that contain the words "Kafka" and "born." Given the vast number of Web pages, Etzioni says, the search engine should be able to notice errors such as one page saying Kafka's birthplace is Peking is less likely to be correct, for example, than the tens of thousands that say Prague.

It's more challenging for a computer to extract more subjective data from text, such as judgments about hotels or movies, but a well-designed algorithm can figure out cues, such as which de-

PHOTOGRAPH COURTESY OF THE UNIVERSITY OF WASHINGTON

scriptive phrase is stronger: clean, almost spotless, or sparkling. It should be able to distinguish the positive—"The room was nice and quiet"—from the negative—"I was disappointed the room wasn't quieter."

Blog and Twitter Searches

One growing area that poses new challenges for search engines is social media, such as blogs, Twitter feeds, and Facebook status updates. "I don't think we have really good blog search yet," says Marti A. Hearst, a professor in the University of California, Berkeley School of Information. Along with Microsoft researchers Susan T. Dumais and Matthew Hurst, Hearst says blog search should be able to accomplish three tasks: find out what people are thinking about a certain topic over time; suggest blogs that are good to read for their style, personality, and other criteria; and find useful information in older blog posts, along the lines of standard search of more static documents.

Blog search needs to take into account the differences between blogs and traditional documents, such as the former's use of more informal language, their different link topology, the importance of timeliness, and the fact that updates tend to not be full HTML pages. Blog search must also take into account that much of the information on blogs is subjective.

To accomplish these tasks, search engine designers look for representations of features that might belong to a particular class of posting, such as the readability level of a page. Machine learning algorithms can then figure out that particular distributions of features may be characteristic of a certain class.

"If you have labeled data and examples of things that you think have a particular attribute, then you can use that to find something similar," says Dumais, principal researcher in Microsoft Research's Adaptive Systems and Interaction Group. But rating postings as positive or negative, or figuring out whether they're aimed at an older or younger audience or have a left-leaning, right-leaning, or middle-of-the-road viewpoint, is challenging, she says. "They do involve a richer understanding of language than most search engines have," Dumais notes.

Search can be improved through a deeper understanding of a document's meaning and a better grasp of a searcher's intentions.

Twitter use has grown explosively in recent months, and in October the company made a deal to open its data to Microsoft's search engine. Dumais says that, with its 140-character limit leading to creative abbreviations of words and condensed hyperlinks, searching Twitter will pose some interesting challenges. But once those are tackled, Twitter users should be able to conduct more refined searches than the service currently allows, while the flow of Twitter data provides search designers with new information that may make search richer. "The volume of the content [on the Web] is actually very useful for some types of algorithms," Dumais says.

One useful fact is that people with Twitter feeds and Facebook pages are making public a lot of information about themselves that search engines can use to better understand their search queries. Just as search can be improved through a deeper understanding of what documents mean, it can also improve through a better grasp of the searcher's intentions. "The real issue with a search engine is not just to serve up results, but to help people accomplish what they're trying to do," says Jon Kleinberg, a professor of computer science at Cornell University.

Search engines trying to provide the right answer to a query might take into account what a user has previously searched for. If a user is looking for a restaurant or a movie recommendation, the search engine might look at the user's friends lists and see what those presumably trusted sources liked. And if the user is searching from a mobile device, that might provide additional clues.

If nothing else, a search from a mobile phone tells the search engine it is from a phone, so perhaps a search for a person is really a search for their phone number. And many mobile devices use GPS or cell phone towers to determine their location. A person typing "Yankees" in Manhattan may be looking for tickets to tonight's baseball game, whereas the same search in Seattle may represent a desire for last night's score. "In a relatively short time frame, we're going to think of geolocation as an integral component of a lot of the online activity we do," Kleinberg says.

Time is also becoming a characteristic to take into account, Kleinberg says. One way of judging the importance of a news story, for instance, is how quickly it spread and how long interest focused on it. Dumais points out that many facts have a time component as well. The gross national product of Norway, the population of Brazil, and the prime minister of Japan—all can have one factual answer in 2000 and a different one in 2010.

Dumais says future search engines will have both a better grasp of the intent of a query and a richer understanding of Web content. "We're looking at how we can support that in ways that go beyond 2.3 words typed into a search box and 10 blue links," she says.

Further Reading

Etzioni, O., Banko, M., Soderland, S., Weld, D. Open information extraction from the Web. *Commun. of the ACM 51*, 12, 2008.

Hearst, M.A. *Emerging Trends in Search Interfaces.* Cambridge University Press, New York, NY, 2009.

Backstrom, L., Kleinberg, J., Kumar, R., Novak, J. Spatial variation in search engine queries. *Proc. 17th Int'l Conf. on World Wide Web*, 2008.

Hearst, M.A., Hurst, M., Dumais, S.T. What should blog search look like? *Proc. of the 2008 ACM Workshop on Search in Social Media*, 2008.

Downey, D., Dumais, S.T., Liebling, D., Horvitz, E. Understanding the relationship between searchers' queries and information goals. *Proc. 17th ACM Conf. on Information and Knowledge Management*, 2008.

Neil Savage is a science and technology writer based in Lowell, MA.