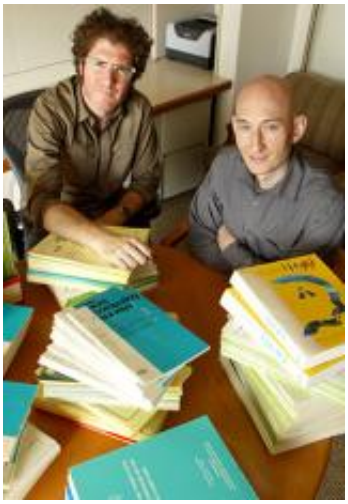
New search method tracks down influential ideas

October 20 2010, by Chris Emery



Princeton computer scientists David Blei (left) and Sean Garrish have developed a new method to search academic journals and other collections of documents, such as websites, to trace the origins and spread of ideas. (Photo by Frank Wojciechowski)

(PhysOrg.com) -- Princeton computer scientists have developed a new way of tracing the origins and spread of ideas, a technique that could make it easier to gauge the influence of notable scholarly papers, buzz-generating news stories and other information sources.

The method relies on computer algorithms to analyze how language morphs over time within a group of documents -- whether they are research papers on quantum physics or blog posts about politics -- and to

determine which documents were the most influential.

"The point is being able to manage the explosion of information made possible by computers and the Internet," said David Blei, an assistant professor of computer science at Princeton and the lead researcher on the project. "We're trying to make sense of how concepts move around. Maybe you want to know who coined a certain term like 'quark,' or search old news stories to find out where the first 1960s antiwar protest took place."

Blei said the new search technique might one day be used by historians, political scientists and other scholars to study how ideas arise and spread.

While search engines such as Google and Bing help people sort through the haystack of information on the Web, their results are based on a complex mix of criteria, some of which -- such as number of links and visitor traffic -- may not fully reflect the influence of a document.

Scholarly journals traditionally quantify the impact of a paper by measuring how often it is cited by other papers, but other collections of documents, such as newspapers, patent claims and blog posts, provide no such means of measuring their influence.

Instead of focusing on citations, Blei and Sean Gerrish, a Princeton doctoral student in computer science, developed a statistical model that allows computers to analyze the actual text of documents to see how the language changes over time. Influential documents in a field will establish new concepts and terms that change the patterns of words and phrases used in later works.

"There might be a paper that introduces the laser, for instance, which is then mentioned in subsequent articles," Gerrish said. "The premise is that one article introduces the language that will be adopted and used in

the future."

Previous methods developed by the researchers for tracking how language changes accounted for how a group of documents influenced a subsequent group of documents, but were unable to isolate the influence of individual documents. For instance, those models can analyze all the papers in a certain science journal one year and follow the influence they had on the papers in the journal the following year, but they could not say if a certain paper introduced groundbreaking ideas.

To address this, Blei and Garrish developed their algorithm to recognize the contribution of individual papers and used it to analyze several decades of reports published in three science journals: Nature, the Proceedings of the National Academy of Sciences and the Association for Computational Linguistics Anthology. Because they were working with scientific journals, they could compare their results with the citation counts of the papers, the traditional method of measuring scholarly impact.

They found that their results agreed with citation-based impact about 40 percent of the time. In some cases, they discovered papers that had a strong influence on the language of science, but were not often cited. In other cases, they found that papers that were cited frequently did not have much impact on the language used in a field.

They found no citations, for instance, for an influential column published in Nature in 1972 that correctly predicted an expanded role of the National Science Foundation in funding graduate science education.

On the other hand, their model gave a low influence score to a highly cited article on a new linguistics research database that was published in 1993 in the Association for Computational Linguistics Anthology. "That paper introduced a very important resource, but did not present

paradigm-changing ideas," Blei said. "Consequently, our language-based approach could not correctly identify its impact."

Blei said their model was not meant as a replacement for citation counts but as an alternative method for measuring influence that might be extended to finding influential news stories, websites, and legal and historical documents.

"We are also exploring the idea that you can find patterns in how language changes over time," he said. "Once you've identified the shapes of those patterns, you might be able to recognize something important as it develops, to predict the next big idea before it's gotten big."

The researchers presented their new method at the International Conference on Machine Learning held this June in Haifa, Israel.

Provided by Princeton University