Researchers develop tools to access 'scholarly big data'

January 28 2014, by Stephanie Koons

Academic researchers and corporate managers often seek experts or collaborators in a particular field to enhance their knowledge or maximize the talents of their workforce. Harnessing that data, however, can be a challenge. Researchers at Penn State's College of Information Sciences and Technology (IST) and the Department of Computer Science and Engineering (CSE) have devised recommendation systems for expert and collaborator discovery that enable users to access "scholarly big data."

"Everyone talks about big data," said Hung-Hsuan Chen, one of the researchers on the project. "But in academia, not many groups have this volume of data from CiteSeer. For data-driven research, we have a very good opportunity for big data research because we are one of the few groups that have such a large volume of data."

Chen, who received a doctorate degree from the Department of CSE in December 2013, was one of the two graduate students from Penn State among the 25 selected from over 250 applicants worldwide to present his research, "CSSeer: an expert recommendation system based on CiteSeerX," at Amazon's first annual Ph.D. Symposium on "Building Scalable Systems" on Nov. 20, 2013 at Amazon headquarters in Seattle. He was supervised by C. Lee Giles, David Reese Professor of IST and Graduate Professor of CSE. In addition to Chen and Giles, the paper was co-written by Pucktada Treeratpituk, computer scientist at the Ministry of Science and Technology in the Thai government; and Prasenjit Mitra, associate professor at the College of IST.

In the paper, the researchers propose CSSeer, a free and publicly available keyphrase based recommendation system for expert discovery based on the CiteSeerX digital library and Wikipedia as an auxiliary resource. CSSeer generates keyphrases from the title and the abstract of each document in CiteSeerX. Those keyphrases are then used to infer the authors' expertise.

"The system automatically figures out who are the experts of a given area," Chen said.

In recent years, the big data phenomenon has been a growing movement with a wide range of social and technological implications. Big data is the term for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications. Big data represents new opportunities for organizations to gather, store, and analyze large volumes of digital information to optimize performance, customer service, and other critical operations. While big data has taken off in the public and private sectors, Giles said, academics generally don't have access to large volumes of digital data that would aid their research. The CiteSeer system was designed to provide them with tools for accessing that information.

"We have access to scholarly big data so we can do experiments on what's 'in' in scholarly big data," he said. "What are the trends, what are people doing research on now, what methods are popular?"

Giles, who directs the Intelligent Systems Research Laboratory, has been involved in the creation and development of various novel search engines and digital libraries. He was one of the creators of the popular computer and information science search engine, CiteSeer, an autonomous citation indexing search engine and digital library. Recently, it has been replaced by the next generation CiteSeer, CiteSeerX- an evolving scientific

literature digital library and search engine that has focused primarily on the literature in computer and information science. CiteSeerx aims to improve the dissemination of scientific literature and to provide improvements in functionality, usability, availability, cost, comprehensiveness, efficiency, and timeliness in the access of scientific and scholarly knowledge. The system, which includes about 3 million documents, gets about 2 to 4 million hits a day.

"It's heavily used by graduate students, professors and researchers throughout the world," Giles said. "It's comparable to an open-source Google Scholar (a web search engine that indexes the full text of scholarly literature across an array of publishing formats and disciplines)."

To put their theories into practice, Giles, Chen and their collaborators recently shipped the CSSeer system to the Dow Chemical Co., an American multinational chemical corporation headquartered in Midland, Mich. that manufactures plastics, chemicals, and agricultural products. According to Giles and Chen, Dow managers and researchers wanted to identify the areas of expertise of some individuals in their organization. The researchers complied with the company's request by building an expert recommendation system for Dow based upon its internal documents.

"If (Dow managers) want to start a project, they can use the system to figure out who are the internal experts," Chen said.

The CiteSeerX system is flexible and "independent of the field," Giles said. Rather than creating just another digital library, CiteSeerx attempts to provide resources such as algorithms, data, metadata, services, techniques, and software that can be used to promote other digital libraries. CiteSeerx has developed new methods and algorithms to index PostScript and PDF research articles on the Web.

Collaborative research has been increasingly popular and important in academic circles, since collaboration brings more points of view to the issues addressed. However, the design of traditional digital libraries and search engines focuses on discovering relevant documents rather than people who share similar research interests. In addition to helping researchers identify experts in a given area, the CiteSeerX platform can also be used to help scholars and scientists discover potential collaborators. CollabSeer is a search engine for discovering potential collaborators for a given author or researcher. The system discovers collaborators based on the structure of the co-author network and a user's research interests.

"Similar to Facebook, (CollabSeer) will recommend individuals based on your social network and key phrases," Giles said.

The long-term plans for CiteSeer, Giles said, include making the system easier and more efficient to run, improving data extraction methods, increasing the number of documents, and automating the extraction of tables and figures in documents. Currently, he added, CiteSeer has over 3 million documents, and "we think we will be able to expand way beyond that."

"We will continue to build upon this large and growing collection of free, publicly available documents," Giles said.

Provided by Pennsylvania State University