New search engine ranks tables by title, document content, text reference

August 8 2007

Penn State researchers have developed a search engine-TableSeer-which not only can identify and extract tables from PDF documents but also can index and rank the search results using factors including the table's title, text references to the table and date of publication.

The engine's innovative ranking algorithm, TableRank, also can identify tables found in frequently cited documents and weigh that factor as well in the search results, said Prasenjit Mitra, an assistant professor in the Penn State College of Information Sciences and Technology (IST) and one of the lead researchers in the development of the search engine.

"TableSeer makes it easier for scientists and scholars to find and access the important information presented in tables, and as far as we know, is the first search engine for tables," Mitra said.

Tables are an important data resource for researchers. In a search of 10,000 documents from s and conferences, the researchers found that more than 70 percent of papers in chemistry, biology and computer science included tables. Furthermore, most of those documents had multiple tables.

But while some software can identify and extract tables from text, existing software cannot search for tables across documents. That means scientists and scholars must manually browse documents in order to find tables-a time-consuming and cumbersome process.

TableSeer automates that process and captures data not only within the table but also in tables' titles and footnotes. In addition, it enables column-name-based search so that a user can search for a particular column in a table.

In tests with documents from the Royal Society of Chemistry, TableSeer correctly identified and retrieved 93.5 percent of tables created in text-based formats, Mitra said.

Searching for tables has some unique challenges, as there is no standard table representation, so tables can appear in PDF, PowerPoint, HTML and Microsoft Word documents. The researchers chose to focus on PDF documents because of their growing popularity in digital libraries and because PDF documents had been overlooked in other table-search efforts.

"Tables can be made using a number of editor tools, and the techniques we are using in TableSeer should work with any text-based tool," said C. Lee Giles, professor of information sciences and technology and co-director of the IST Cyber-Infrastructure Lab where the research originated. "While we designed and developed TableSeer to facilitate searching of tables occurring in articles in the chemistry domain, it can be used in any domain where data is presented in tabular form including other scientific, technical, social and business areas."

The development of TableSeer is part of an open-source cyber-infrastructure project focusing on chemical document search for environmental chemistry and funded by the National Science Foundation. The grant awarded to the Penn State Department of Chemistry aims to enable automatic data analysis.

"Searching and extracting information from data tables is an essential component of data analysis in environmental science, where many

research groups publish large amounts of kinetic data describing chemical changes in the environment," said Karl Mueller, professor of chemistry and principal investigator for the NSF grant.

"As we approach multidisciplinary problems within the Penn State Center for Environmental Kinetics Analysis, our students spend many days hunting down and compiling large amount of data from tables. The TableSeer tools will definitely increase the efficiency of this process and allow more time to be spent on creative scientific analysis," he added.

TableSeer can be tested online (see chemxseer.ist.psu.edu). The source code will be made available near the completion of the project, the researchers said.

In the meantime, research is ongoing to improve the ranking algorithm by adding additional features. The researchers also are working on a search engine that can identify, extract and rank figures found in documents, as figures are another important device for disseminating data and findings in the natural sciences.

Source: Penn State