New text-mining algorithm to prioritize research on chemicals, disease for public database

April 17 2013

Keeping up with current scientific literature is a daunting task, considering that hundreds to thousands of papers are published each day. Now researchers from North Carolina State University have developed a computer program to help them evaluate and rank scientific articles in their field.

The researchers use a text-mining algorithm to prioritize research papers to read and include in their Comparative Toxicogenomics Database (CTD), a public database that manually curates and codes data from the scientific literature describing how environmental chemicals interact with genes to affect human health.

"Over 33,000 scientific papers have been published on heavy metal toxicity alone, going as far back as 1926," explains Dr. Allan Peter Davis, a biocuration project manager for CTD at NC State who worked on the project and co-lead author of an article on the work. "We simply can't read and code them all. And, with the help of this new algorithm, we don't have to."

To help select the most relevant papers for inclusion in the CTD, Thomas Wiegers, a research bioinformatician at NC State and the other co-lead author of the report, developed a sophisticated algorithm as part of a text-mining process. The application evaluates the text from thousands of papers and assigns a relevancy score to each document.

"The score ranks the set of articles to help separate the wheat from the chaff, so to speak," Wiegers says.

But how good is the algorithm at determining the best papers? To test that, the researchers text-mined 15,000 articles and sent a representative sample to their team of biocurators to manually read and evaluate on their own, blind to the computer's score. "The results were impressive," Davis says. The biocurators concurred with the algorithm 85 percent of the time with respect to the highest-scored papers.

Using the algorithm to rank papers allowed biocurators to focus on the most relevant papers, increasing productivity by 27 percent and novel data content by 100 percent. "It's a tremendous time-saving step," Davis explains. "With this we can allocate our resources much more effectively by having the team focus on the most informative papers."

There are always outliers in these types of experiments: occasions where the algorithm assigns a very high score to an article that a human biocurator quickly dismisses as irrelevant. The team that looked at those outliers was often able to see a pattern as to why the algorithm mistakenly identified a paper as important. "Now, we can go back and tweak the algorithm to account for this and fine-tune the system," Wiegers says.

"We're not at the point yet where a computer can read and extract all the relevant data on its own," Davis concludes, "but having this text-mining process to direct us toward the most informative articles is a huge first step."

More information: Davis AP, Wiegers TC, Johnson RJ, Lay JM, Lennon-Hopkins K, et al. (2013) Text Mining Effectively Scores and Ranks the Literature for Improving Chemical-Gene-Disease Curation at the Comparative Toxicogenomics Database. *PLOS ONE* 8(4): e58201.

doi:10.1371/journal.pone.0058201