Researchers create search engine to hunt molecules online

July 26 2007

ChemxSeer, the first publicly available search engine designed specifically for chemical formulae, can sort out when "He" refers to helium and not a person more than nine times out of 10, according to the Penn State College of Information Sciences and Technology (IST) researchers who created the tool.

With the new engine, scientists searching for research on CH4 or methane no longer have to wade through search results about Channel 4 or Chapter 4 as ChemxSeer will only return documents with references to the chemical formula.

The new algorithm also can identify related chemicals with different formula representations and chemicals with related substructures or similarities, said C. Lee Giles, professor of information sciences and technology and co-director of the IST Cyber Infrastructure Lab where the research originated.

"Results from our search engine are much more relevant than results returned by popular search engines," Giles said. "It is one of several cyber tools under development in our lab which will enable better access to and sharing of information and data among scientists and scholars."

The tool is described in a paper, "Extraction and Search of Chemical Formulae in Text Documents on the Web," presented at the recent 16th International World Wide Web Conference in Alberta, Canada. In addition to Giles, the authors are Bingjun Sun and Qingzhao Tan,

graduate students in computer science and engineering, and Prasenjit Mitra, assistant professor of information sciences and technology and co-director of Penn State's Cyber Infrastructure Lab.

Electronically hunting for chemical formulae poses some unique challenges for popular search engines, which typically focus on key words. For one, scientists often search for parts of chemical formulae, with the part appearing in the beginning, at the end or in between.

Similarly, some chemical molecules can have more than one formula representation. As a result, if a person is searching for CH4 using a popular search engine and the article identifies the molecule as H4C, the article won't be included in the search results.

In addition, molecules can be confused with nonchemical abbreviations. While people would recognize "OH" as Ohio in a particular context, a machine with a chemical dictionary could confuse it with the chemical notation for a hydroxide. A similar slip up can occur with "I" (iodine) or "In" (indium).

In designing the engine, the researchers built on their expertise in information-extraction algorithms created for CiteSeer, a search engine for academic and science documents.

Besides extracting formulae, ChemxSeer also allows for various query models appropriate for any scientist looking for a molecule. Not only does it query for exact matches, but it also queries for formulae with additional terms or elements as well as for formulae with similar structures. The engine also can search for the range of occurrence of an element in various formulae, the researchers said.

To create ChemxSeer, the researchers basically "taught" machines how to recognize chemical formulae by providing training samples of

occurrences of both chemical formulae and non-chemical formulae.

"Teaching the computer to classify what is a formula and what is not was complex because language is inherently context sensitive and judging the meaning of a term using its context is hard for a machine," Mitra said.

Future research will focus on improving the reliability of identification, linking to existing molecular databases, data archiving and increasing the relevance of search results.

The engine 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.

"This tool replaces time-intensive manual searching, allowing our research team to focus more on solving problems with as much relevant information as possible," said Karl Mueller, professor of chemistry and PI of the cyber infrastructure grant.

Source: Penn State