

OCTOBER 04, 2004

Textpresso's Richer Blend of Scientific Data

Frustrated that the volume and increasing complexity of the scientific literature might make it impossible for researchers to keep pace, Howard Hughes Medical Institute researchers have developed Textpresso, a new text-mining system that sifts through the scientific literature and identifies relevant information nearly as well as human document curators.

Textpresso was originally developed by HHMI researchers to organize a database of 7,000 research articles and 20,000 scientific abstracts about research on the roundworm, *C. elegans*. The scientists are now making the software available to the scientific community to organize the rapidly growing databases of scientific literature on other organisms and disciplines.

Textpresso's developers, Hans-Michael Müller, Eimear Kenny, and Howard Hughes Medical Institute investigator Paul W. Sternberg, all of the California Institute of Technology, published a report on the project in the November 2004 issue of the journal, *Public Library of Science Biology*. The article was published online ahead of the print publication.

"I find that I have more time to read papers, because the computer is skimming the papers for me to find the ones I really need to read. I didnt expect that to happen."

- Paul W. Sternberg

Sternberg said Textpresso was born from a sense of frustration that he and his colleagues encountered in organizing WormBase, a collection of scientific data on the roundworm. "One problem was that we had these very highly trained biologists trying to extract information from papers and put it into a uniform, accessible format that people can access," said Sternberg. "It's slow and tedious, and we wanted to speed it up and automate it—especially to keep these talented, smart, motivated people happy."

“The other issue was that we wanted to give biologists a more efficient text-mining tool to enable them to find the information they needed more efficiently,” he said.

Sternberg said that the enormous increase in complexity of scientific literature with the advent of large-scale genomics projects makes the need for advanced text-retrieval systems critical. “Every year I’ve been doubling the amount of papers that I personally needed to read,” he said. “Now it’s hundred-fold worse than just a few years ago.”

In contrast to search engines that rely on individual keywords, Textpresso uses an ontology to organize information in a text database, said Sternberg. The Textpresso ontology consists of a catalog of types of objects and concepts for which the sentences in articles can be searched. These objects and concepts include terms like “gene” or “cell,” but also terms that relate objects, such as “association” and “regulation,” or describe an object, such as “biological process.”

Once the ontology was created for roundworm literature, the Textpresso program automatically found those terms in the text database and labeled them using the document markup language XML.

After the documents have been coded, users can search the database in a far more sophisticated and useful way than can be done using just keywords, said Sternberg. Searches can consist of more meaningful semantic queries that use combinations of terms in the ontology to extract more precise results.

“When you type in keywords, you’ll certainly find a lot of material you want, but also a lot you don’t,” he said. “But the kind of categorization in Textpresso enables searches with greater precision, yielding more relevant information and less irrelevant information,” he said.

In evaluating Textpresso’s accuracy, the researchers compared how Textpresso and human curators performed in identifying sentences in a set of journal articles that contained information on the interaction between two particular genes. Textpresso identified about 62 percent of the sentences and the human curators found about 71 percent.

“In general, we found that there were things that humans missed and things that the machine missed,” said Sternberg. However, he said, the performance was comparable between the two. In larger tests, Textpresso was used to search for information on interaction between two particular genes. The researchers found that Textpresso searched with far greater efficiency.

Sternberg and his colleagues are now making Textpresso available for organizing and searching other databases, including the scientific literature on *Drosophila*, yeast, and neuroscience topics. Adapting Textpresso involves first “stripping down” the ontology to categories that are general to all the

databases. Then objects and concepts particular to the database can be added to the ontology, said Sternberg. To enable full interdisciplinary use of the system, the researchers also will need to change the underlying database structure. The team believes the system can be extended to automatically extract facts that can then be placed in a database such as WormBase with only minor oversight from an expert curator.

Sternberg said that commercial journals should find the system advantageous because a Textpresso search would yield only individual sentences relevant to the search. A scientist would then obtain the full article through the regular channels. Conversely, scientists could use Textpresso to determine whether particular papers are relevant before accessing them through a subscription.

Personally, said Sternberg, Textpresso has made his quest for scientific information far more efficient. "I find that I have more time to read papers, because the computer is skimming the papers for me to find the ones I really need to read, and I didn't expect that to happen."