Stimulus grant will improve physics arXiv

November 18 2009, By Bill Steele

(PhysOrg.com) -- Stimulus funding will enhance Cornell's e-print arXiv of scientific papers to help users identify a work's main concepts, see research reports in context and easily find related work.

"It shouldn't be a one-way channel," said Paul Ginsparg, professor of physics and information science, who heads the new project funded by a three-year $883,000 grant from the National Science Foundation, with federal stimulus money from the American Recovery and Reinvestment Act (ARRA).

The arXiv currently contains close to 600,000 papers in physics, mathematics, computer science, quantitative biology, quantitative finance and statistics, with some 5,000 new papers submitted each month. Researchers submit their work as "preprints" before formal publication. Such preprints used to be passed around by hand before Ginsparg launched the arXiv (pronounced "archive") in 1991 at the Los Alamos National Laboratory; he brought it to Cornell in 2001, where it is now hosted by Cornell Library.

New tools will link papers by concepts, not just by the citations they contain, and this will help users without advanced expertise -- including some outside the scientific community -- understand the significance of new research, said Ginsparg.

"One of the underlying concepts of the arXiv was leveling the playing field," he explained. "Formerly, new research was available only to a few privileged people. Now everyone has equal access, but because of

differential levels of expertise not [all scientists] can as easily assess significance. We will be working on automated tools to help identify and highlight the most important concepts," he said. Along with scientists, he added, the site is closely watched by journalists.

The system also will identify related databases and commentaries. For example, Ginsparg said, if a paper mentions an astronomical object, the computer could serve up a menu of related information, including a database describing the object, the original observations that generated the description, and blogspace commentary.

Computers usually search documents by looking for specific words or phrases, but concepts are not always described with the same exact words, and some words mean different things in different places. New algorithms will use a "fuzzier" approach, inferring concepts by the ways terms are used, and will track related documents over a five- or 10-year time scale, so users will be able to see the "genealogy" of ideas. Newer documents will be linked to such data as definitions and rules for reasoning about it, which enables machines to infer relationships.

Other enhancements will provide interoperability with such research sites as PubMedCentral and provisions to allow scientists to contribute in newer, more flexible text formats.

Researchers might be more enthusiastic about participating in open access journals and repositories if they could see that their work was more accessible and usable, Ginsparg suggested. "And perhaps the academic community will again play a role at the forefront as the semantic Web 3.0 rolls out," he said. Academic publishing has lagged behind the commercial Internet in providing interactive enhancements that today's students take for granted, he explained. "Configuring research communications infrastructure for the next generation of researchers requires getting into the heads of near-term future

researchers -- undergrads and grad students -- coming of age in the Google/Facebook/Twitter era."

The project is expected to generate jobs for two graduate students and a half-time programmer. To date, Cornell has received 124 ARRA grants, totaling more than $99 million.

Provided by Cornell University (news : web)