Vertical search across the educational horizon

December 22 2010

Searching the web usually involves typing keywords or a phrase into a search engine and clicking the "search now" button. It's very effective and several large companies have become prominent in the field by providing users with searchable access to millions, if not billions of web pages in this way. However, according to researchers at Hewlett Packard in Palo Alto, California and Chinese technology company, Innovation Works, general search engines, while very effective at tracking down information, are nevertheless unstructured, which limits the user's ability to further automate the processing of the search results.

Other researchers have attempted to find ways to support more precise web searching on specific sites, so-called content verticals, but writing in the *International Journal of* Computational Science *and Engineering*, HP's Meichun Hsu and IW's Yuhong Xiong explain an alternative web search system that could be used to search across such verticals. They have demonstrated how the new system works by focusing on online courses.

The researchers point out that in the pre-web days, a relational database within a company or educational establishment was equivalent to the modern online content vertical. Users of relational databases could embed their search results in an application program for that database. The HP team hopes to take forward this embedding process and extend it to the wider web. As an example of the kind of search such an approach might allow they describe how they would like to be able to carry out the following:

SELECT product_name FROM hp.com WHERE product_type PC

Imagine how a similar query across online educational resources might be made transparent to users by clever programming so that they could pull up specific prospectuses, curricula, timetables, and tests quickly and easily, across domains rather than on a single computer system. To solve this problem the team has exploited "focused crawling" in which only the pages likely to be relevant are crawled and indexed. This ties in neatly with "web content classification", which adds meta-data to those relevant pages that accelerates searching. Finally, "information extraction" pulls out the important information from that focused and classified data. The team has now applied this approach to HP's OfCourse project.

"The technologies can be used to support structured queries over contents extracted and aggregated from the web," the team says. "They are also foundational to personalization, by offering more insights into the web content of interest to particular users." The new approach to search does require human intervention at certain stages so that contents within each domain crawled might be classified more effectively, but machine learning approaches can also lead to some degree of automation of this process too. The research, the team says, takes us one step closer to "the convergence of database technology and information retrieval in the era of the web."

More information: "Scalable information extraction for web queries" in *International Journal of Computational Science and Engineering*, 2010, 5, 176-184

Provided by Inderscience Publishers