Culture vultures go beyond, way beyond Google

December 22 2008

(PhysOrg.com) -- European researchers are pushing online culture and heritage research way beyond Google by using a smart search system that is multilingual, multimedia and optimised for cultural heritage. Better yet, this promising system has wide application in other fields.

European researchers have developed an optimised search system that can access an enormous quantity of cultural heritage resources that reside online. Current technology like Google takes a scattergun approach, dishing up dozens of links of sometimes variable quality.

"Right now, if you do a search online, you get lots of irrelevant overload," explains Pasquale Savino, coordinator of the MultiMatch project, which set out to create state-of-the-art search technology for cultural heritage information.

The MultiMatch system targets searches using a variety of smart search methods. Better yet, the concept can be applied to other fields, like sport, politics, economics and technology.

"Consider that many portals already offer a specialised catalogue, but in many cases the selection and classification of data is done manually, while the MultiMatch platform can perform this work automatically," Savino reveals.

Three trumps

Savino says MultiMatch trumps standard search in three, vital ways. "The system does not simply query the web, it also searches through archives, many of them not publically available," he notes.

Archives like the National Library of Austria (ONB), Biblioteca Virtual Miguel de Cervantes and the Israel National Library, though currently the system accesses just a portion of these resources for research purposes.

It also supports multimedia searches, and not simply by looking for pictures by name. It can look for pictures using other pictures. If a user has one picture, say of Picasso's Guernica, the system can search for images in a similar style. It can do the same types of search for sound and video resources, too.

MultiMatch is also fluent in six languages. A search entered in Polish can be targeted to look for results in Spanish, or English, Italian, Dutch and German – the other four languages that the system currently recognises.

Finally, MultiMatch presents its results in an aggregated way, with resources clearly identified by type and sorted by priority, whether it is

relevance, historical period or some other criteria. It is a prioritised, sorted and easily grasped layout of results, a bit like a newspaper created on the fly, for your particular query.

Culture crawl

MultiMatch began by selecting well-known cultural heritage sites like the Biblioteca Virtual Miguel de Cervantes, to populate its database. Next, it used well-known cultural heritage websites to 'train' web crawlers.

A web crawler is an automated program that accesses a website and traverses through the site by following the links present on the pages. Crawlers index links and information found on the various websites.

The MultiMatch crawlers are self-learning, so after they were shown cultural heritage websites, followed by sites that were not related to cultural heritage, the crawlers 'learned' what to look for. Over time, the system becomes self-refining, as it learns appropriate and inappropriate websites.

The system can also identify relevant material via an in-depth 'crawling' of selected cultural heritage institutions. And the system is not just multilingual, it speaks metadata as well, the lingua franc of the 'semantic web' – an attempt to help machines 'understand' the context and significance of specific types of data.

The result is that MultiMatch can take advantage of whatever metadata descriptions are in place, typically in an archive.

But MultiMatch goes further. If there is no metadata, it tries to infer the semantic content of a page – what it means and what it refers to – and this, too, is self-learning, and so will improve over time.

MultiMatch can also automatically extract information which can then be used to create cross-referencing, via hyperlinks, between related material, such as the biography of an artist, exhibitions of his/her work, a video documentary or critical appraisals, and so on.

Obsessive wikis

"We hope, in the future, to take functionality further, so that you could search for Cubism, for example, or any art movement. The query would return a categorised and prioritised table of contents for that very specific topic. The system can not do that yet, but it is something we want to develop in the future," Savino explains.

It would be like a personalised Wikipedia, created on the fly, that caters uniquely to your obsession with Cubism.

In the meantime, Savino and the MultiMatch team are focusing on three prototype demonstrators to test the prototype of the system. One will support teachers trying to develop a lesson plan, the other two will focus on archiving and tourism applications that are still to be finalised.

Commercial breaks

The team, however, do not expect any surprises in the tests: the system has been working reliably in the lab up to now. Once it is validated, however, several of the partners will incorporate aspects of the work into their commercial products.

"There is also the possibility that one of the partners will develop a new product from our results," notes Savino, though he emphasises that the current platform version is a prototype and would need more work to make into a commercial product.

The lead partner, Savino's ISTI-CNR, will keep the demonstrator running online for at least one year after the project finishes, in the winter of 2008, to carry on further work.

In the meantime, MultiMatch technology will be used in two other European-funded projects: Europeana, a major effort to provide online access to 2 million digital objects from the continent's archives, museums and libraries, and the European Film Gateway, a similar project specialising its work in moving images.

"These projects are mainly using the technology we developed to ensure interoperability between different archiving systems, and multilingual search and discovery," confirms Savino.

It is an indication of the value of the Multimatch search technology.

The MultiMatch project received funding from the ICT strand of Sixth Framework Programme for research. The technology was also showcased at the ICT 2008 meeting in Lyon.

MultiMatch project: www.multimatch.eu/

Provided by ICT Results