Reaping Deep Web Rewards Is a Matter of Semantics

Greg Goth

Case-sensitive passwords are common on the Internet. And now, perhaps, the global community of Internet developers might have to prepare for a case-sensitive semantic Web.

The lowercase *s* semantic Web that might offer users a far richer Internet experience differs dramatically from the Semantic Web — a specific framework defined by the W3C — that has hovered tantalizingly just out of mass reach for almost a decade. Researchers are exploring several different approaches to providing an alternate semantic experience, from domain-specific academic research engines to commercial offerings from established companies such as Google and startups such as Kosmix (www.kosmix.com).

One thing these researchers have in common, however, is a belief that the top-down approach that orthodox Semantic Web technologies use isn't something the world should wait for.

Google research scientist Alon Halevy says the company's researchers hew to a "bottom-up" approach of divining ways to supply users with richer and wider results, a diametrically opposed approach to that of the Semantic Web.

"They are starting in a top-down way, which is to say, 'Let's create a language for representing knowledge and relationships between different pieces of data, and then let's get people to annotate their data in some way according to the languages we design.' There's a chicken-and-egg problem there."

The problem Halevy refers to is the fact that no contemporary mass search technology can search annotated content; hence, developers have no incentive to painstakingly create the annotations that would make their content more visible. He's seconded by Anand Rajaraman, cofounder of Kosmix, a startup engine that gives users a composite page pulled from other sources and classified using an algorithmic taxonomy. In some sense, this approach fulfills the general concept of providing a semantically useful Web page — the taxonomy algorithm can infer matches between search terms such as "Catalina Island" and travel advertisements and reviews that focus on the California coastal island, or "Gettysburg," which returns pages on the pivotal battle of the American Civil War and the liberal arts college located in the Pennsylvania town.

"We've been waiting for the uppercase Semantic Web since 1995 or so," Rajaraman says. "It's been a very long wait, and I don't think it's going to get any shorter. The lowercase *s* semantic Web is the way the world is going to evolve. There is absolutely no incentive for people to apply semantics to their data, but there is every incentive for companies like Kosmix and others to use semantics from data that's already there."

Researchers Go Deep

Numerous academic and commercial researchers are exploring ways to access and index the contents in the deep Web — mainly content hidden behind HTML forms in databases — in order to offer users more information to answer their queries. Currently, much of the academic research centers on domain-specific aspects of the deep Web because academic funding is inadequate to tackle form discovery and indexing over the entire Web. The communities might approach the issue from slightly different vectors, but there is near consensus that the crux of delivering richer material to the Internet user is developing a way to access deep Web and surface Web pages and provide some sort of semantically aware architecture to constrain query results.

"When structured data in the deep Web is surfaced, the structure and hence the semantics of the data is lost," Halevy and coauthors of "Harnessing the Deep Web: Present and Future" (www-db.cs.wisc.edu/cidr/cidr2009/Paper_115.pdf) wrote. "The loss in semantics is also a lost opportunity for query answering."

For example, the authors presented a theoretical search for a particular brand, year, and model of automobile — "used ford focus 1993."

They then posed the following scenario: there is a surfaced used-car listing page for Honda Civics, which has a 1993 Honda Civic for sale, but with a remark "has better mileage than the Ford Focus."

The problem with such a return, the authors say, is that the infor-

1089-7801/09/$25.00 © 2009 IEEE

News in Brief

As directed by the recently passed **American Recovery and Reinvestment Act**, the US **Federal Communications Commission** in April officially launched its effort to create a comprehensive national strategy for achieving affordable and **universal broadband access**. In its notice of inquiry, the FCC asked for public comment on the effectiveness of both existing market conditions and government policies and regulations, as well as on the best steps forward for improving and expanding the nation's digital infrastructure. The stimulus bill funnels more than **US$7 billion for network build-outs** through commerce and agriculture department agencies; the FCC is charged with developing an overall broadband plan and delivering it to Congress in February 2010.

More information is available at http://hraunfoss.fcc.gov/edocs_public/attachmatch/DOC-289900A1.pdf.

The **ITU Telecommunication Standardization Sector's Focus Group on ICTs and Climate Change** announced that it has developed a **method for calculating energy usage and carbon impact** from ICT lifecycles, as well as the decrease in greenhouse gas emissions that ICTs can achieve through efforts such as **"dematerialization"** which replaces atoms with bits (for example, by buying an MP3 file rather than a CD). The group, which includes some of the world's leading ICT players, agreed on four deliverables in a March meeting in Hiroshima, Japan. Those deliverables draw on best practices from international organizations and will be published as proceedings. The ITU will hold its third **Symposium on ICTs and Climate Change** in Quito, Ecuador, on 8–10 July 2009.

More information is available at www.itu.int/themes/climate.

According to a recent **Verizon Business Security Solutions** report, criminal organizations were responsible for more than 90 percent of the **285 million electronic record exposures** in 2008. The report also found that most of those exposures could have been avoided by applying basic preventative measures.

More information is available at www.verizonbusiness.com/products/security/risk/databreach.

The ninth annual **French Big Brother Awards** ceremony was held on 4 April and — in a move that reflected what organizers call "the increase of surveillance and social control in France" — officially honored almost one-third of its nominees. Among the "winners" was **Michèle Alliot-Marie, the French interior minister**, who was given the organization's Lifetime Menace award for her "immoderate taste for police files" and her talent for constructing the "internal enemy." The awards, sponsored by the French chapter of **Privacy International**, included many other dubious Orwell honors as well as one positive one: the annual Voltaire award. This year, that award was shared among several groups, including a coalition united against Base élèves — the central database of information on children — and another coalition fighting the use of biometrics in schools.

More information is available at http://bigbrotherawards.eu.org/Les-decorations-promotion-2009.html.

A group of organizations led by the Washington-based nonprofit **Public Knowledge** is calling on **President Obama** to broaden his appointments to **intellectual property positions** to include **diverse IP stakeholders**. Several recent appointees represent industries that support broad IP protection, including the recording industry. According to Public Knowledge, such industries "might favor established distribution models at the expense of technological innovators, creative artists, writers, musicians, filmmakers, and an increasingly participatory public."

More information is available at www.publicknowledge.org/node/2070.

mation-retrieval style index of the search engine might consider that a good result. However, the semantics of the user's query are unknown. The user might want to buy a 1993 Ford Focus, find parts from a 1993 Ford Focus, or might already own one and needs some maintenance information. The fact that a 1993 Honda gets better gas mileage isn't relevant to that user's needs.

"Such a scenario could be avoided if the surfaced page had the annotation that the page was for used-car listings of Honda Civics and the search engine were able to exploit such annotations," the authors concluded. "Hence, the challenge here is to find the right kind of annotation that can be used by the IR-style index most effectively."

There is an inherent irony in tackling this issue, however, as noted by James Geller, professor of computer science at the New Jersey Institute of Technology, who served as chair of "The Semantic Web Meets the Deep Web" workshop held in conjunction with the 2008 IEEE Joint Conference on E-Commerce Technology and Enterprise Computing. Geller's colleague, Soon Ae Chun, professor of information systems at the City University of New York's College of Staten Island, served on the workshop's program committee.

"The insight that motivated us to start our research on the topic was that it is theoretically much easier to extract knowledge from a well-structured table compared to free English text," Geller and Chun wrote in response to questions from *IEEE Internet Computing*. "Many researchers have been trying to build ontologies automatically by extracting concepts from online documents."

However, Geller and Chun assert that such efforts rely on an assumption of sophisticated natural language processing yet to be reached.

"On the other hand, if we can observe that a column in a table, hidden

behind a Web front end, contains the data items Mexico, Spain, France, Japan, etc., then we can make a pretty good guess, as humans, that all terms in this column describe countries. If we can then extract all of them automatically and include them in an ontology, then our program has just learned a lot of useful information about the world."

However, Nancy Ide, chairman of the computer science department at Vassar College, says the different levels of sophistication in data mining and computational linguistics don't offer much near-term promise.

"From the point of view of computational linguistics, people have done the Gigaword corpus, and we don't yet have the technology to get real semantic meaning out of large data sets just by doing statistical analysis," Ide says. "It's the old 80-20 rule. We can disambiguate words 80 percent of the time, but it's the hard part we'll never get. So, having a lot more data is not the big answer."

Many Small Answers Under Way

If, as Ide says, having a lot more data isn't the big answer to providing semantically enriched results, trying to find ways to provide relevant resources in discrete pieces of that data continues in the hope that many small steps will advance the larger goal. One recently launched deep Web resource, DeepPeep (www.deeppeep.org), is being constructed under the guidance of Juliana Freire, associate professor of computer science at the University of Utah. DeepPeep currently offers access to 13,000 Web forms in seven domains. It contains elements of Freire's research such as a hidden-Web crawler that automatically retrieves data behind keyword-based interfaces; a focused crawler that efficiently locates sparse concepts on the Web, and which the Utah researchers used to locate online Web databases and services; a classifier ensemble that can determine the domain of Web forms with high accuracy; a clustering strategy for organizing a large set of Web forms; and a learning-based approach for automatically extracting labels from Web forms.

Freire says her research team is working on issues such as figuring out which part of a page of structured data contains useful information "because you have navigation bars, ads, and lots of junk. So, we are working on how to detect templates and remove the junk."

Geller and Chun wish to go beyond indexing-form labels and form-field values of the deep Web.

"The DeepPeep search engine looks for domain-specific forms that may lead users to desired deep Web contents," they say. "Our initial approach to extracting Web form labels, to use them as index terms, is similar to their approach reported in VLDB [Very Large Data Base] 2008. However, what we advocate is to annotate the forms in a way such that even the generic search engines such as Google and Yahoo can locate the deep Web forms. This requires not only the labels used in the forms to be indexed, which seems to be the predominant method used in DeepPeep, but [that] the semantic contents of the deep Web also be available for search."

Next Steps

Significant advances in semantic enrichment — in both the deep and surface Web — will face different hurdles in different settings. Although academic researchers might be hampered by the inability to create an infrastructure scalable enough to attract large numbers of users, commercial entities such as the large search engines might find it difficult to alter their revenue-producing platform architectures to accommodate nascent semantic technologies, even technologies that don't rely on ortho-

dox Semantic Web elements such as the Resource Description Framework and Web Ontology Language.

Kosmix's Rajaraman, who in 1996 created the pioneering comparative e-commerce shopping engine Junglee with Venky Harinarayan (who is also a Kosmix cofounder), is confident the company's taxonomy-and-topic approach might give users the presentational semantics of cross-referenced data and documents on a page that doesn't demand unprofitable labor to create or deliver disappointing query returns.

"At Junglee, we did some deep Web work around data from shopping and classified ad sites and so on," he says, "so we've been doing this for a very long time. This time it's not just shopping, it's everything."

And, Rajaraman says, success of such lowercase semantic technology could render the Semantic Web moot for mass audiences.

"There might be some specific niche areas where the capital *S* Semantic Web might happen, especially in vertical industries where a few dominant players can dictate structure," he says. "But I don't think its going to happen in areas where there is fragmentation. The small *s* semantic Web is what's going to happen, and we'll realize the uppercase Semantic Web is not necessary after all."

Greg Goth is a freelance technology writer based in Connecticut.