The Dark Side of the Semantic Web

Intelligent Readers,

I’ve recently been giving a talk with the same name as this column at various locales. I originally hoped that the provocative title would catch peoples’ eye and might indeed raise eyebrows, as I’m obviously not known as a critic of the Semantic Web. On the contrary, I remain a committed tech-evangelist for this important new technology. I’m not using “dark side” as in Darth Vader and the Dark Side of the Force, but rather as in “dark side of the moon.” The point is, a lot is happening with the Semantic Web that’s exciting and important that comes from the “Web” side, rather than the AI side. So, to many AI researchers this part of the technology is unknown—thus, the “dark side” allusion.

To understand this trend, which some call “Web 3.0” (perhaps unfortunately), we need to understand the Web not as a linked set of documents but as a technical construct of protocols, processes, languages, and tools that make it all work. While there’s no way to go into all that in this short editorial, we can gain some insight by looking at other emerging trends on the Web and examining how a little bit of AI can have a big effect.

Learning from Web 2.0

One problem with the Web’s ubiquity and importance is that it’s often hard to tell the marketing from the meat for various Web applications. The trend that came to be known as “Web 2.0” (again, perhaps unfortunately) started from a fairly specific core of technologies (essentially Web services and Ajax) but became the name by which almost everything new and exciting on the Web came to be known. Wikipedia, Flickr, the “blogosphere,” social-networking sites, and YouTube, to name just a few, have been new Web applications associated with this so-called next-generation Web. (Tim Berners-Lee, as you can read in his book Weaving the Web, included such applications in his original Web vision, so they might more aptly be considered the realization of the original “Web 1.0” and not a new generation of technology. But such is the marketing needed to make things happen in Silicon Valley.)

From an AI viewpoint, the most interesting thing about Web 2.0 applications has been the use of tagging to associate keywords with nontext items. Photo and video sites have taken great advantage of this approach, as have social-bookmarking approaches (such as del.icio.us) and various players in the Wiki and blogging space.

“Ahh,” cried the critics, “The Semantic Web is overkill. Folksonomies and social processes are all we need to make this work.” Perhaps they were right—that is, up to a certain point. And that point is being reached now. In retrospect, it seems obvious to many (and to many in the AI community it was obvious from the beginning) that this technology can scale only to a certain level. Here’s a simple thought experiment. Suppose we tag every photograph in Flickr with all the text needed to capture every concept in each photo—the “thousand words” that the picture is worth. Then take the tens of millions of these photo documents and ask how to search for them on the basis of these keywords. Sounds a lot like what Google was created for, doesn’t it? So how would these unstructured, undisambiguated, nonsemantically aligned keywords somehow create, as many of their advocates claimed, a naturally occurring semantics that would somehow challenge the rule of the keyword-based search engine? Up to a certain size, statistics look great and work well (and clustering was a key to early Web 2.0 successes). But beyond a certain size, making statistical retrieval work for language is nontrivial (and a mainstay of many AI’s human-language-technology researchers for whom the taggers’ claims never held water). In short, the taggers are learning one of AI’s recurring themes—that which looks easy in the small is often much harder in the large.
A little semantics

For many AI researchers, this social part of the Web really is like the dark side of the moon. We’re so used to thinking that “knowledge is power” that we fall into a slippery-slope, more-is-better fallacy. If some expressivity is good, lots must be great, and in some cases this is correct. What we forget, however, is something that’s become sort of a catch phrase in Semantic Web circles. “A little semantics goes a long way.” In fact, I’m just now beginning to understand exactly how little is needed to go a long way on something as mind-bogglingly huge and unorganized as the Web.

A key realization that Berners-Lee had with respect to RDF’s design is that having unique names for different terms, with a social convention for precisely differentiating them, could in and of itself be an important addition to the Web. If you and I decide to use “http://www.cs.rpi.edu/~hendler/elephant” to designate some particular entity, then it really doesn’t matter what the other “blind men” think it is. They won’t be confused when they use the natural language term “elephant,” which isn’t even close, lexicographically, to the longer term you and I are using. And if they choose to use their own URI, “http://www.other.blind.guys.org/elephant,” it won’t get confused with ours.

The trick comes, of course, as we try to make these things more interoperable. It would be nice if someone from outside could figure out and, even better, assert in some machine-readable way that these two URIs really designate the same thing—or different things, or different parts of the same thing, or … Oops, notice how quickly we’re on that slippery slope. If we want all the ways we could talk about how these things relate, we’re back to rediscovering knowledge representation in all its glory (and the morass of reasoning issues that come with). But what if we stop somewhere and allow only a little bit of this? Suppose we simply go with “same” or “different.” Sounds pretty boring and not at all precise—certainly not something that will get you an article in IEEE Intelligent Systems.

Ah, but now let’s move to the Web. Whenever a user creates a blog entry in livejournal.com, a small machine-readable description becomes available against a somewhat minimal person ontology called FOAF (friend of a friend). I read recently that livejournal has about 15 million FOAF entries. Other blogging and social-networking sites also create FOAF files, accounting for at least 50 million little machine-readable Web documents. And FOAF contains a small little piece of OWL (the Web Ontology Language), which says we should assume that two entries having the same email address are by the same person. This one little piece of “same” information suddenly allows a lot of interoperability and a jump start for many kinds of data mining applications. The rule might not be 100-percent correct, but then on the Web, what is?

Believe it or not, several start-up companies are successfully using this rule, and other similar assertions about equality (or inequality), to help create personal-information-management tools for Web data or to better match advertising to Web users. (I hate to say it, but such matching is the biggest legal moneymaker on the Web.) By being able to, even heuristically, equate things found in different Web applications, a whole range of mash-ups and other Web applications become possible. A very little piece of semantics, multiplied by the billions of things it can be applied to on the Web, can be a lot of power.

Semantic raisins

There’s more to this dark-side story that gets into technical aspects of Web application development. Forthcoming Semantic Web standards, such as the SPARQL query language or the GRDDL (Gleaning Resource Descriptions from Dialects of Languages) mechanism for adding semantic annotations to XHTML pages, make it much easier to embed these little bits of semantics into other Web applications (including those of Web 2.0). So while the funding from places such as DARPA and the Nation Science Foundation in the US, and from the EU’s Information Societies Technology program, has been looking at what we might call the Semantic Web’s “high end,” the leading edge of the Web world has been bumping into the “low end” and finding useful solutions available in the RDF and OWL world.

Expert systems never really made it as a stand-alone technology, but they were far more successful when appropriately embedded in other applications (Pat Winston’s...
“raisins in the raisin bread”). Semantic Web developers are beginning to understand that our technology can similarly gain use by being successfully embedded into the somewhat chaotic, but always exciting, world of Web applications. This opens up a brand-new playground for us to explore largely unexamined approaches in which a little AI, coupled with the very “long tail” of the Web, opens up exciting possibilities for a very different class of (just a little bit) intelligent systems.

Welcome to the Dark Side,

James Herdian