Could Google Become the Semantic Web?

The Semantic Web, a far-reaching, complex new standard described as a "Web of data" is not too helpful for most people, I have learned. There's a lot of talk about it, but for most Web users, the idea described in 1981 by Tim Berners-Lee means little.

A number of experts and forward thinking companies do care though, including Google. The patent applications I reviewed while waiting to give a lecture on the topic in 2007 were authored by Ramanathan Guha. The key document, in my opinion, was filed in April 2005, published on February 15, 2007, as "Programmable Search Engine", US2007 00386616. (The mathematical procedures are explained in US2007/00386616, US2007/0038601, US2007/0038603, US2007/0038600, and US2007/0038614. Additional information may be found in the published literature related to the semantic Web and in Google's collection of technical papers in the Google Labs' subsite.)

As an advisor to BearStearns & Co. at that time, I reported in an email that Google had a far-reaching invention in the Ramanathan Guha filings. Furthermore, Google filed on the same day a total of five patent applications related to what Google called the "Programmable Search Engine." My analysis of these documents revealed a solid anchoring in the functions in what I call the "semantic space."

BearStearns published a report containing my analysis in May 2007 as part of equity research for its consumer Internet division. The report caused a minor ripple in the financial world and the information did not reach the popular press due to the limited circulation these dense, technical equity reports get.

Dr. Guha and Tim Bray of OpenText worked on the Meta Content Framework, and Google was making strategic and competitive use of some of the ideas in the W3C spell out Semantic Web activity. What was interesting was that five other "semantic" inventions were filed at about the same time by Google's attorneys and each was authored by the luminary by Dr. Guha, who had worked at IBM Almaden, founded Epinions.com, and worked on the W3C MTF spell out project.

In 2006, Google acquired a company founded by Dr. Alon Halevy, a respected researcher in what most people describe as data management. The term is misleading because based on my research, Dr. Halevy's work has implications well beyond what most experts would put under the heading of "database research."

Dr. Halevy's work complemented Dr. Guha's and since 2006, the two strands of research have become part of a broader semantic capability that Google continues to develop at this time. (My analysis of Dr. Halevy's work is available from IDC as Report #213562, published in September 2009.)

Simplifying Dr. Halevy's brilliant work, let me say that it makes possible different types of queries, using some of the features of Dr. Guha's work and new methods that go well beyond Dr. Halevy's patent documents filed when he worked at units of the Bell Labs and Lucent's

research arm. The idea is to enrich an information object with additional tags so that queries about lineage (where something came from) and likelihood of accuracy (the "correctness" of an information element) can be used to generate a result.

When I read the patent applications in February 2007, I realized that Google was operating at a high level in the field of semantic analysis. Its systems and methods were, in my opinion, much more sophisticated than I had realized. Simplifying again, my research supported my argument that Google was developing techniques that could figure out the meaning of content and "fill in the blanks" when an item of information was ambiguous or missing. Not surprisingly, Google was investing in numerical recipes to enhance its existing systems' software that "reads and understands" the meaning of discourse.

The patent applications included several interesting features. For example, people with information - such as a Web site operator - could "push" information to Google. The same technique appears in Google's patent documents related to indexing video content, for instance. The idea is that humans can "teach" Google's system certain things.

A second important concept is the context server. The context server provides a knowledgebase that other Google methods can access for the purposes of determining the metadata required to determine the conceptual meaning of a system process, user action, information object, or any other object processed by the Google system. In simple terms, a context server holds information Google has discovered about a topic, person, company, etc "for future reference" so that Google's system can disambiguate or "fill in the blanks" when additional information about a document is needed.

A third component, which gains significance now that we have public versions of Google Wave and Google Buzz to explore, is that explicit user actions can be used to index certain content objects. Examples range from clicks on hyperlinks to changes made by users to certain content objects within the Wave.

One of the key concepts is the idea of context. For example, when a zoologist searches for "bat", results should be different from the ones shown to a 4th grade student. The PSE spell out can review previous queries, past clicks, and time spent on a page (called "dwell time" by Google) to gauge the intent of a user's search - once again this evokes the concept of Google becoming a Database of Intentions, described in John Battelle's book *The Search*. Further, another important concept is that of Google leveraging the Web - Webmasters (site owners) will be the ones creating the files telling Google what to index, how to index, and what is allowed. Moreover, the PSE will "learn" as it digests metadata provided by the millions of Webmasters.

The heart of the intelligence, therefore, is "context files" with information about users, data, Web sites, and programmatic processes that execute under certain conditions. It is also noteworthy that while site owners can have a lot of control over this process, Dr. Guha's invention includes processes that reduce Google's dependence on webmasters' actions. The PSE can in certain circumstances create the various XML files on its own.

Implications

The semantic initiatives at Google have significant implications for such competitors as IBM, Microsoft, Oracle, and Yahoo, and many other companies working in content processing, content aggregations, and text mining. Google's methods disclosed in its open source documents may make the content in social systems such as Facebook and Twitter outputs more intelligible, although that remains to be demonstrated in Google's new initiatives in social search.

In my opinion, Google's capabilities in semantic methods could give the firm a competitive advantage in certain types of content processing.

Other implications of Google's semantic methods include:

- Smarter services for enterprise customers, including value added indexing and autonomous software agents that operate on metadata provided by the context server
- More granular detail in search results; for example, car inventory on a lot in addition to the local dealer's phone number and location on a map
- Better spam filtering
- Access to "deep Web" or "invisible Web" content
- Metatagging of non text information (audio and video content)
- Cross generated content from different sources (a dossier on a person such as Michael Jackson prepared by an algorithm, not a human writer for a traditional publication)
- More sophisticated ad matching for text and rich media

The semantic Web is a logical evolution of content available via Web sites. One of the members of my research team asked, "Could Google become the Semantic Web?"

My instinct was to reject the question as specious. After working through Google's technical papers in preparation for my forthcoming study on Google's non-text indexing methods, I am not so sure. Google's share of the Web search market continues to creep upwards. Depending on whose data I examine, Google's share is between 65 and 75 percent. Google's seeping into other market sectors such as telecommunications and education yields significant reach beyond the browser-based search model. New service demonstrations such as Google Squared and the structured query embedded in "normal" Google search results are rich with significance. Run the query "SFO LGA" from Google.com and examine the results. You can see Google's system figuring out the query, creating a parametric search, and giving you one-click access to travel listings. Although a small demonstration, Google's semantic power is humming under the clean Google interface.

There are implications for intelligence, national security, and public policies activities as well.

Net Net

Google, in my view, is a key player in the Semantic Web. As Google *becomes* the Internet for many users, it may be that Google's methods define the Semantic Web by virtue of its market presence.

At least one member of my team believes that Google *is becoming* the Semantic Web. I am not yet convinced, but I am tracking Google's open source information in this important field of information science. Others may want to focus their lasers on this facet of Google as well.

Stephen E Arnold, ArnoldIT.com at www.arnoldit.com/sitemap.html

Mr. Arnold is a consultant residing in Harrod's Creek, Kentucky. You can learn more about Google in his three studies of Google technology available from Infonortics, Ltd. in Tetbury, Glos., UK: *The Google Legacy* (2005), *Google Version 2.0* (2007), and *Google: The Digital Gutenberg* (2009). His most recent Google monograph will be published by Ovum, part of the Datamonitor Group, in the United Kingdom in mid 2010. You can follow Mr. Arnold's public observations in the Beyond Search Web log at http://www.arnoldit.com/.wordpress and in the Strategic Social Networking blog at http://ssnblog.com.