

November 12, 2006

Entrepreneurs See a Web Guided by Common Sense

By [JOHN MARKOFF](#)

SAN FRANCISCO, Nov. 11 — From the billions of documents that form the World Wide Web and the links that weave them together, computer scientists and a growing collection of start-up companies are finding new ways to mine human intelligence.

Their goal is to add a layer of meaning on top of the existing Web that would make it less of a catalog and more of a guide — and even provide the foundation for systems that can reason in a human fashion. That level of artificial intelligence, with machines doing the thinking instead of simply following commands, has eluded researchers for more than half a century.

Referred to as Web 3.0, the effort is in its infancy, and the very idea has given rise to skeptics who have called it an unobtainable vision. But the underlying technologies are rapidly gaining adherents, at big companies like [I.B.M.](#) and [Google](#) as well as small ones. Their projects often center on simple, practical uses, from producing vacation recommendations to predicting the next hit song.

But in the future, more powerful systems could act as personal advisers in areas as diverse as financial planning, with an intelligent system mapping out a retirement plan for a couple, for instance, or educational consulting, with the Web helping a high school student identify the right college.

The projects aimed at creating Web 3.0 all take advantage of increasingly powerful computers that can quickly and completely scour the Web.

“I call it the World Wide Database,” said Nova Spivack, the founder of a start-up firm whose technology detects relationships between nuggets of information mining the World Wide Web. “We are going from a Web of connected documents to a Web of connected data.”

Web 2.0, which describes the ability to seamlessly connect applications (like geographical mapping) and services (like photo-sharing) over the Internet, has in recent months become the focus of dot-com-style hype in Silicon Valley. But commercial interest in Web 3.0 — or the “semantic Web,” for the idea of adding meaning — is only now emerging.

The classic example of the Web 2.0 era is the “mash-up” — for example, connecting a rental-housing Web site with Google Maps to create a new, more useful service that automatically shows the location of each rental listing.

In contrast, the Holy Grail for developers of the semantic Web is to build a system that can give a

reasonable and complete response to a simple question like: “I’m looking for a warm place to vacation and I have a budget of \$3,000. Oh, and I have an 11-year-old child.”

Under today’s system, such a query can lead to hours of sifting — through lists of flights, hotel, car rentals — and the options are often at odds with one another. Under Web 3.0, the same search would ideally call up a complete vacation package that was planned as meticulously as if it had been assembled by a human travel agent.

How such systems will be built, and how soon they will begin providing meaningful answers, is now a matter of vigorous debate both among academic researchers and commercial technologists. Some are focused on creating a vast new structure to supplant the existing Web; others are developing pragmatic tools that extract meaning from the existing Web.

But all agree that if such systems emerge, they will instantly become more commercially valuable than today’s search engines, which return thousands or even millions of documents but as a rule do not answer questions directly.

Underscoring the potential of mining human knowledge is an extraordinarily profitable example: the basic technology that made Google possible, known as “Page Rank,” systematically exploits human knowledge and decisions about what is significant to order search results. (It interprets a link from one page to another as a “vote,” but votes cast by pages considered popular are weighted more heavily.)

Today researchers are pushing further. Mr. Spivack’s company, Radar Networks, for example, is one of several working to exploit the content of social computing sites, which allow users to collaborate in gathering and adding their thoughts to a wide array of content, from travel to movies.

Radar’s technology is based on a next-generation database system that stores associations, such as one person’s relationship to another (colleague, friend, brother), rather than specific items like text or numbers.

One example that hints at the potential of such systems is KnowItAll, a project by a group of [University of Washington](#) faculty members and students that has been financed by Google. One sample system created using the technology is Opine, which is designed to extract and aggregate user-posted information from product and review sites.

One demonstration project focusing on hotels “understands” concepts like room temperature, bed comfort and hotel price, and can distinguish between concepts like “great,” “almost great” and “mostly O.K.” to provide useful direct answers. Whereas today’s travel recommendation sites force people to weed through long lists of comments and observations left by others, the Web. 3.0 system would weigh and rank all of the comments and find, by cognitive deduction, just the right hotel for a particular user.

“The system will know that spotless is better than clean,” said Oren Etzioni, an artificial-intelligence researcher at the University of Washington who is a leader of the project. “There is the growing realization

that text on the Web is a tremendous resource.”

In its current state, the Web is often described as being in the Lego phase, with all of its different parts capable of connecting to one another. Those who envision the next phase, Web 3.0, see it as an era when machines will start to do seemingly intelligent things.

Researchers and entrepreneurs say that while it is unlikely that there will be complete artificial-intelligence systems any time soon, if ever, the content of the Web is already growing more intelligent. Smart Webcams watch for intruders, while Web-based e-mail programs recognize dates and locations. Such programs, the researchers say, may signal the impending birth of Web 3.0.

“It’s a hot topic, and people haven’t realized this spooky thing about how much they are depending on A.I.,” said W. Daniel Hillis, a veteran artificial-intelligence researcher who founded Metaweb Technologies here last year.

Like Radar Networks, Metaweb is still not publicly describing what its service or product will be, though the company’s Web site states that Metaweb intends to “build a better infrastructure for the Web.”

“It is pretty clear that human knowledge is out there and more exposed to machines than it ever was before,” Mr. Hillis said.

Both Radar Networks and Metaweb have their roots in part in technology development done originally for the military and intelligence agencies. Early research financed by the [National Security Agency](#), the [Central Intelligence Agency](#) and the Defense Advanced Research Projects Agency predated a pioneering call for a semantic Web made in 1999 by Tim Berners-Lee, the creator of the World Wide Web a decade earlier.

These agencies also helped underwrite the work of Doug Lenat, a computer scientist whose company, Cycorp of Austin, Tex., sells systems and services to the government and large corporations. For the last quarter-century Mr. Lenat has labored on an artificial-intelligence system named Cyc that he claimed would some day be able to answer questions posed in spoken or written language — and to reason.

Cyc was originally built by entering millions of common-sense facts that the computer system would “learn.” But in a lecture given at Google earlier this year, Mr. Lenat said, Cyc is now learning by mining the World Wide Web — a process that is part of how Web 3.0 is being built.

During his talk, he implied that Cyc is now capable of answering a sophisticated natural-language query like: “Which American city would be most vulnerable to an anthrax attack during summer?”

Separately, I.B.M. researchers say they are now routinely using a digital snapshot of the six billion documents that make up the non-pornographic World Wide Web to do survey research and answer questions for corporate customers on diverse topics, such as market research and corporate branding.

Daniel Gruhl, a staff scientist at I.B.M.’s Almaden Research Center in San Jose, Calif., said the data

mining system, known as Web Fountain, has been used to determine the attitudes of young people on death for an insurance company and was able to choose between the terms “utility computing” and “grid computing,” for an I.B.M. branding effort.

“It turned out that only geeks liked the term ‘grid computing,’ ” he said.

I.B.M. has used the system to do market research for television networks on the popularity of shows by mining a popular online community site, he said. Additionally, by mining the “buzz” on college music Web sites, the researchers were able to predict songs that would hit the top of the pop charts in the next two weeks — a capability more impressive than today’s market research predictions.

There is debate over whether systems like Cyc will be the driving force behind Web 3.0 intelligence will emerge in a more organic fashion, from technologies that systematically extract meaning from the existing Web. Those in the latter camp say they see early examples in services like del.icio.us and Flickr, the bookmarking and photo-sharing systems acquired by [Yahoo](#), and Digg, a news service that relies on aggregating the opinions of readers to find stories of interest.

In Flickr, for example, users “tag” photos, making it simple to identify images in ways that have eluded scientists in the past.

“With Flickr you can find images that a computer could never find,” said Prabhakar Raghavan, head of research at Yahoo. “Something that defied us for 50 years suddenly became trivial. It wouldn’t have become trivial without the Web.”

[Copyright 2006 The New York Times Company](#)

[Privacy Policy](#) | [Search](#) | [Corrections](#) | [RSS](#) | [First Look](#) | [Help](#) | [Contact Us](#) | [Work for Us](#) | [Site Map](#)
