Steven Arnold | The search continues

GCN Interview with Steven Arnold of the Google Government Report

By Joab Jackson, GCN Staff

"I really believe FirstGov does much more focused indexing [than Google]." Steven Arnold

Steven Arnold got an early start on search engines. In 1971, his employer, Halliburton Co., assigned him to digitize the company's technical reports in order to make them searchable. He has worked in the field ever since. In the past decade, he has moved over to consultancy, starting his own practice, Arnold IT. In 2000, he helped generate the technical plan for the first iteration of the General Services Administration's FirstGov government search engine. (His son, Erik Arnold, currently works on FirstGov.) More recently, he launched the Google Government Report (www.ggreport.com), a newsletter and electronic information service offering tips on how to be better recognized by Google. We caught up with Arnold to get his views on what is happening with both enterprise and Web search.

GCN: What is the state of search these days?

Steven Arnold: This has been a time when people are realizing that enterprise search doesn't work. Folks with enterprise search systems are really on the lookout for technologies that make search more useful for the users.

So what will carry us into 2007 is a collection of technologies we think of as text mining, where software algorithms look at documents and find the names of people, places and things and attempt to relate them to one another.

GCN: What is wrong with search technologies today?

Arnold: Take a real-life example: You and your significant other go to England and she says on the way home, "I loved that jumper that I saw at Harrods."

So, if you're like me you don't have any clue about what she's talking about. So now I go to a search system, like at Neiman Marcus Online or Overstock.com, and type in the word "jumper." What I get back is not that sweater that she saw in Harrods. I get stuff back unrelated to that idea. And that is a very common problem.

GCN: What are the text mining companies doing that the search companies can't?

Arnold: From a technical point of view, companies like Attensity Corp. and nStein Technologies are not focused on search. They are focused on figuring out the nuances, relationships and the important concepts in a document. Their systems generate index terms that an enterprise search system can suck in.

GCN: So how does this technology work?

GCN.com | Most Read Articles | Most E-Mailed Articles

Past 24 hours | Last 7 Days | Last 30 Days

- Seven policies to watch in 2007
- Florida taps ACS for KidCare deal
- Seven programs to watch in 2007
- Census to modernize data distribution systems
- Nevada selects state CIO

Go to complete list

Arnold: Every one of the new companies that I have looked at—and I have a text mining report where I have been tracking this—are approaching the problem both mathematically and by doing vocabulary and knowledge-based analysis. Their software decomposes sentences into subjects, verbs and adjectives and analyzes the results with the predictive algorithms.

What is unusual is that the computer chips are so darn powerful now. These new companies are basically saying that they are going to use these chips and throw everything we learned in computer science at the search problem and it will work out just fine.

So what you have now are hybrid [search/text mining] systems and by golly they are very interesting. When you use one of those [services] against scientific and technical information, your accuracy can be 85 or 90 percent. If you run it against general business news, you can hit 80 to 85 percent accuracy. That's almost as good as a human search.

The system can run automatically, and when something [unusual] comes up, you have a person who has been through school look at the error report and fix the mistakes. When someone spells Al Qaida differently, and an exception comes up, an analyst can look at that and say, "This is an OK spelling."

GCN: So we had no idea this sea change was happening in the industry.

Arnold: Yeah, it is going on under everybody's nose because so much focus is given to Google.

Google actually has some nifty technology like this, but it is so anchored in the consumer space. These smaller companies are like cigarette boats racing around the destroyer.

The age of innovation for this is not over. I've been at conferences this year where people are saying, "Yeah, well it's over. Microsoft is giving it away, or Oracle is there, Google is there. There is no room for innovation." I just don't agree.

GCN: Why should agencies care about Google?

Arnold: Let me give you an anecdote. I got invited to meet with the people from a large insurance company in Denmark. I asked how much of their traffic came from Google [searches]. And they said they thought about 35 to 40 percent. I asked if there is a way to check, and they said they could do it right there from a laptop. [When they checked], they looked at me and said "You know what? Last month Google was 80 percent of our traffic."

GCN: So you are seeing an increase in Google-derived traffic within the last few months?

Arnold: Literally, within the last 12 to 16 weeks. The anecdote underscores what we've seen in other work, that Google is basically the search engine of choice for virtually everyone in the world.

As people realize how much traffic comes to them from Google, it becomes more important to understand what other people are doing to make sure their Web site is indexed by Google and their sites come up in the context of the proper keywords. You really now have to pay attention to Google not because Google is the greatest company on the planet, but because Yahoo and Microsoft just haven't done that good of a job competing.

GCN: So what can agencies do to better present their pages to Google?

Arnold: The first step is to create a sitemap that conforms to Google's guidelines, because Google has already convinced Microsoft and Yahoo to follow its formula. So that is job one.

Job two is to take a very hard look at the page names and the URLs on your site. Most government Web pages I look at have very long and complicated URLs, and Google's robots can process those, but they prefer to process human-understandable URLs.

The third thing is the government needs to do a better job with content. The government has great information. The Department of Agriculture has outstanding information, but it is presented in such a way that makes it really hard to index and search effectively. If you want a good report, you have to download a huge PDF file.

So I think the government has to make more of its content easy to comprehend, and not put out these 5-megabyte globs of data.

GCN: Any thoughts on the battle between the two premier U.S.-government-focused search engines, FirstGov and Google U.S. Government search?

Arnold: There is no battle at all. I really believe FirstGov does much more focused indexing.

When Google sends its robots to an agency Web site, it looks at the links, indexes the first 100,000 characters per page and follows the links two levels down. But FirstGov looks hard at these sites and goes very deep into the site. Remember the FirstGov [result] set will be much smaller and more focused than the Google set, which will be very broad.

If you work for an information service, you certainly can start a search with Google, but if you want to be thorough, you will have to look at FirstGov. If you're a government worker, you might want to start with FirstGov, but you definitely want to take a look at what Google has indexed.

So think of FirstGov as drilling down into a topic and Google going very broad across many topics. So the two services are complementary.

More news on related topics: Content / Record Management, Management, Web services / XML