How Google Has Changed Enterprise Search

by Stephen E. Arnold
Arnold Information Technologies

In 2004, Google "won" the Web search race — in terms of media coverage, investor excitement, and coining the new verb *google*, as in, "Google it again, Sam." However, though Web searching looms largest in the minds of most knowledge workers, Google's less visible role in enterprise search is becoming as important as Google and Web searching. How does Google's search-and-retrieval technology index and make available an organization's information for intranet users? In recent months, Google's "Appliance" — a search toaster that makes enterprise search much less complicated to implement than other offerings — has been closing sales deals with commercial, nonprofit, and governmental organizations. Google may well be poised to win the enterprise search race as it has the Web search race. Let's look at the impact Google is making upon buyers and sellers of software to make an organization's information searchable.

Dark Matter in the Search Universe

Theoretical physicists talk about "dark matter." In a nutshell, *dark matter* is — according to a Google "hit," of course — "The matter which is postulated to account for at least 90 percent of the mass of the universe, but which has yet to be directly detected. The evidence for its existence comes mainly from observations of the dynamics of stars in galaxies and of galaxies in clusters of galaxies, from gravitational lensing and from cosmological models."

Google is the dark matter of the enterprise search universe (and almost any other type of search). This is one surprising finding from the results of a 6-month study entitled "The Enterprise Search Report," completed in August 2004 for CMSWatch.com. The results of this study of 26 enterprise search systems were published by CMSWatch and are available from the publisher [http://www.cmswatch.com].

For many professionals, the Google software system defines how online search and retrieval should work. Inevitably, Google has become the search system that other vendors have to sell against even if Google sales executives are not pursuing a particular account. Google is simply there, and, like dark matter, it has a significant effect in many search system sales processes. Wherever they go, competitor search system vendors hear the dreaded question, "Is your system like Google's?"

It is common knowledge that Google has millions of users who submit more than 250 million queries per day in dozens of languages. Google's brand has a larger footprint than any other search-and-retrieval system's image. A clean and simple interface, attractive options, different collections of useful content, and unobtrusive commercial messages — these characteristics are secondary to Google's ability to deliver a useful answer. Whether a single term query like *Spears* or a complex "advanced query" specifying PowerPoint presentations by Microsoft's Bill Gates, Google delivers a useful hit almost every time.

In fact, Google has even created a Microsoft-specific search service arguably better than the one provided by Microsoft itself. Check it out at http://www.google.com/microsoft.html. Google's search technology gives Microsoft a fair shake; it lists its own search service for Microsoft lower in results than Microsoft.com. Even when it comes to current or imminent competitors, Google seems to live up to its "Don't be evil" company slogan.

Google Is Search

For our purposes, let's assume that the term "search" means "Google" to most people. It then follows that "google" has become a synonym for search. Only mathematicians and *Jeopardy* winners know the technical definition of a google. For most people, "google" has only one meaning, and that meaning has nothing to do with the word "googol" (a quantity equal to the numeral one followed by 100 zeroes). The mathematical googol sprung from the fertile mind of the American mathematician Edward Kasner in 1938.

With a more upscale demographic than Yahoo!'s user profile, Google has become the dark matter — notice, I did not say *dark force* — in the search universe. And Google's power is not theoretical. For most professionals,
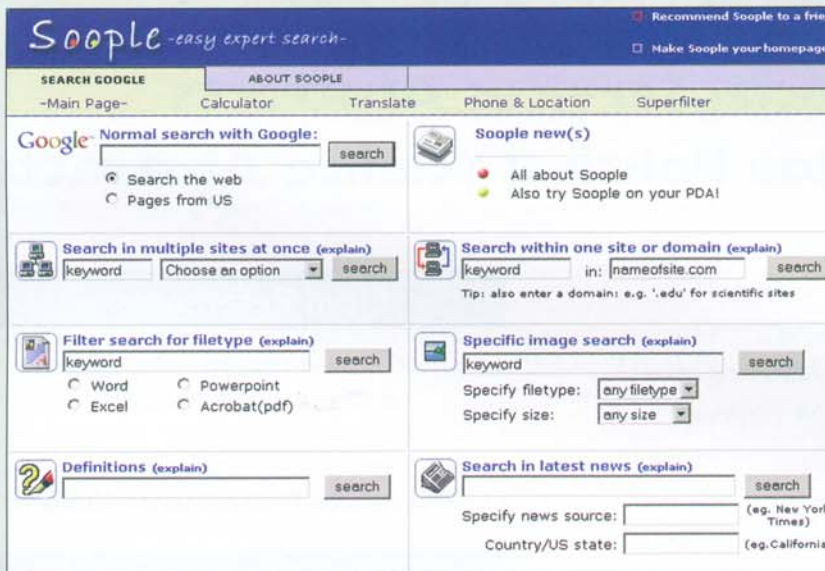

Table 1. Comparison of Public Web Search and Enterprise Search

Public Internet Search (e.g., Google)	Enterprise Search (e.g., Verity K2)
Advertisers pay the cost of search. (Users perceive search as free.)	The organization pays the cost of search. (Users per-ceive the search as free.)
"Smart" algorithms deliver a usable answer virtually all of the time. "Search is magic," says one user.	Delivering the specific piece of information for a busi-ness user requires stored queries, human specialists, and advanced algorithms.
Search results display quickly, often constrained only by the user's connection. Caching and other techniques enhance speed.	Results may take considerably longer than a typical Internet search. Scaling an enterprise search environ-ment is often more complex than organizations first anticipate.
A simple, clean interface or a hybrid interface with point-and-click headings and a search box.	There is no generally accepted way to expose search in an enterprise.
Mass market focus.	Departmental or special need focus. An organization is not a mass market; it is a federation of special needs.
Mostly Internet-accessible content such as HTML, Word files, an Extensible Markup Language (XML). Sometimes others, e.g., PDF.	Multiple file types ranging from standard office soft-ware files to proprietary file types used by proprietary software systems.
Finding and indexing content easily automated.	Finding and indexing content requires additional logic to handle versions, user access rights, and other con-straints such as regulatory requirements.
Spiders handle most content with a small, but increasing amount of FTP or "pushed" content.	A combination of automated indexing, push technology, and proprietary.
Updates scheduled when the Internet indexer decides.	Content must be indexed on a cycle determined by the needs of the business; that is, some content may be indexed in near real time.
Security protects certain aspects of the system; access generally subject to modest security processes.	Security is a critical pivot point. Access, user control lists, hacking prevention, and other concerns are a top priority.
Dedicated team of engineers, usually supplemented with content specialists, form the Internet search com-pany's team.	Dedicated engineers and content specialists are not funded or used on an as-needed basis.
Fail over needs to be "good enough" so users and advertisers don't complain.	Fail-over and redundancy essential but, in many instances, not always provided. A system failure often produces a knee-jerk reaction among the management team.
Subjective results are desirable. Advertisements, paid placements, and in-line advertisements ubiquitous.	Objective results are desirable. Some boosting but usu-ally limited to organizational messages.
Usage tracking is widespread. Results are used to alter the content spidered and the results displayed.	Usage tracking is a complex issue. Tracking is essential for regulatory, security, and copyright requirements. Employee knowledge of tracking is hazy, and it is essential that the enterprise chart a course that meets the needs of the enterprise, regulators, and users. A misstep may lead to litigation or loss of credibility.

applications such as Grokker, a visualization utility for Google hits. Google is creating what Silicon Valley buzzword mavens call an *ecosystem*. The bigger the ecosystem, the harder it is to sign the death warrant on a company. Like Microsoft before it, Google's management feeds the ecosystem with innovations like Gmail with its one gigabyte of storage and an ability to search one's messages. With such innovations, the dark matter becomes momentarily visible and increases what I call "the Google effect."

3. Entrepreneurs and loyal users create what seems like an unending parade of parodies, utilities, and "watch sites." Examples of parodies are scarce because "The Google Trademark Enforcement Team" stamps out spoofs with Microsoftian finality. The spoof at http://www.goegel.be/trademark.html was notified on Aug. 26, 2004, that the site violated the Google trademark (see Figure 1 on page 10). Utilities include Copyscape [http://www.copyscape.com], which provides an easy way for a user to locate a Web page's content in another Web page. Authors have a powerful, free way to find unauthorized copies of their work instantly. Scoople [http://www.scoople.com] is a service that provides a single Web page with search boxes for most Google features, including file type queries, a way to search a single Web site indexed by Google (see Figure 2 on page 10). Scoople makes powerful features accessible to the most inexperienced online user.

Google now thrives in a datasphere — a digital ecosystem in a networked world. Google is important. Like other influential companies, Google has Web sites devoted to monitoring its day-to-day actions. (See, for example, http://www.google-watch.org/.) A side effect of the dark matter is the diffusion of the Google brand and the Google search technology.

Neither Google nor other search engines make clear what they index, when, why, or how thoroughly. Editorial policies are difficult to pin down. The average person looking for infor-

The Google Appliance

Right now, Google is making life difficult for Thunderstone and the more than 100 other vendors of enterprise search products. These companies should not feel singled out. Google is largely unaware of the enterprise search market. The company's senior executives know most of the Big Four in enterprise search (Autonomy Ltd., Convera, Fast Search & Transfer, and Verity Corp.). Many Google engineers know about the innovative work underway at Cambridge University, Carnegie-Mellon University, Syracuse University, and other university research facilities. The smaller and newer companies — for example, Arikus in Toronto, Your Amigo in Australia, and Mondosoft in Copenhagen — are still not on the corporate radar.

Google sells a search appliance. Google in a Box begins at about $35,000 and peaks in the $250,000 range. The interface looks "just like Google on the Web." There are some important differences. First, the price. A Google search on http://www.google.com is free to the user. Costs are paid by advertisers who deposit money with Google. The advertiser has a myriad of choices, but most boil down to renting a keyword such as travel or airline tickets. When a visitor to the Google Web site enters the term `travel` or `airline tickets`, advertisements appear. Advertisers get the nod when it comes to listing Web pages in certain situations.

mation does not care about these niggling details. The thirst is for an answer. Information professionals do care, but they represent only a tiny fraction of the Google user base.

To its credit, Google spews out innovations at a blinding pace, particularly in comparison to the seemingly glacial changes at America Online, CompuServe, Microsoft, and Yahoo. Google is the company most likely to become the Microsoft of 2005. In short, Google has yet to make a serious misstep, and its impact is felt in virtually every situation when search and retrieval are discussed.

How far Google's reach will extend remains unclear. What is clear is that Google is poised to have considerable success in licensing its technology for use inside organizations. Enterprise search is a modest market compared to Google's online advertising revenue. But enterprise search is a market that could balloon to $3 billion or more in less than 36 months.

Enterprise Search

One of the unfortunate aspects of language is that a single word such as *information*, *precision*, or *search* can have many meanings. Enterprise search refers to the use of search and retrieval technology to locate information accessible via an organization's network. Enterprise search is not Web search, although enterprise search can include the indexing of content that resides on Internet sites. Generally, how-

ever, enterprise search is designed to meet the specific search needs of a particular organization and its employees. In some cases, enterprise search will support queries from special classes of authorized users, as in an extranet. For example, General Motors allows dealers and certain vendors to have access to the organization's network for business purposes.

Table 1 on page 12 identifies other important differences between Web search and enterprise search.

Like most corporate activities, the similarity among search engines has an upside and a downside. The upside is that the best-known companies in enterprise search have learned how to market against one another. The companies with the largest market share have emphasized customization, linguistic features, and integration with enterprise software.

The downside is that these companies — BG (before Google) — focused on one another and dismissed the Web search space. One fascinating example comes from the Norwegian company, Fast Search & Transfer. FAST sold its Web search technologies and AlltheWeb.com site to Yahoo! at the time Google was gaining momentum in Web search. FAST's management team shifted from Web search to enterprise search with a highly customizable and concomitantly complex product called ESP. Google, on the other hand, has extended its reach from Web search into the enterprise. Google sells simplicity, an attribute lacking in many enterprise search systems whose complexity can give systems professionals migraines. For FAST's sales professionals, every time Google wins an account, it faces a grim reminder of the short-sightedness of the company's strategy in abandoning the Web search sector. Not surprisingly, Google has captured high-profile accounts and has blindsided the likes of Autonomy, Convera, and Verity.

Google's Web presence is one part of an increasingly effective sales strategy. Unless Google makes a misstep, traditional vendors of enterprise search sys-
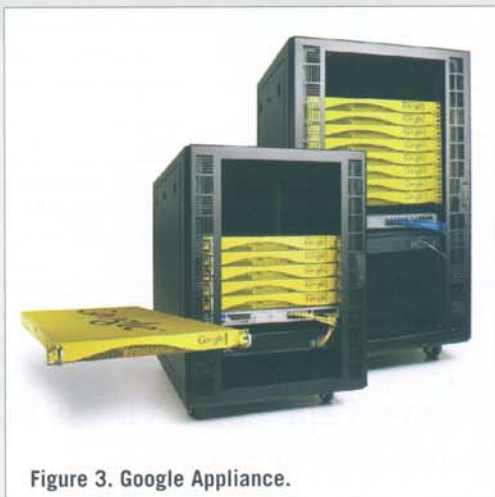

Figure 3. Google Appliance.

tems face a very different type of competitor. Google may not be the information technology department's choice, but it is the user's choice.

In the last 12 months, research for Enterprise Search (published by CMSWatch.com) underscored the pervasiveness of Google. One senior manager at a Fortune 500 company said in an interview, "Search inside our company should work like Google. Instant. Easy to use. Answers on the first page of results. Is that so difficult?" The answer to the manager's rhetorical question is, "Yes, madam. Search like Google's is difficult. Very difficult."

The second difference, then, is objectivity. Google's search system focuses less on the objective search results and more on advertising. The "page ranking algorithm" that determines a particular Web page's relevance based on the number of other Web pages linking to a particular page, among other things, still operates. Judging from the chatter on bulletin boards and listservs where advertisers exchange information, Google's algorithm is tuned, adjusted, or changed depending on factors that outsiders find difficult to penetrate.

Third, Google has now become a public company with revenues that dwarf those of the total revenues of Autonomy, Convera, Fast Search & Transfer, and Verity *combined*.

In a very short span of time, most enterprise search companies have a different financial and brand benchmark against which to present themselves. With regard to money, few enterprise search vendors can match Google's financial position. Instead, search and retrieval technology vendors find themselves having to differentiate their technology from Google's. In astronomy, as in all sciences, one can detect an object in one of two ways: either by observing it directly, or observing the effect that it has on other, more easily observed, objects. Google's effect can be viewed both ways easily.

Google has become synonymous with easy, effective search. Busy customers want a solution, not a lecture on natural language processing. Differentiating a complex search system from Google's search toaster approach becomes a tough sell when potential buyers get confused with discussions of Bayesian statistics, lemmatization, and taxonomies. Unless the vendor is extremely skillful, differentiating one's product from Google can lead to a slippery slope. Endeca and InQuira, two "search companies" which have effectively competed against Google, have found success by shifting the focus from search to guided navigation and seamless integration into specific functions such as customer support. For vendors without a way to outflank Google, the sales cycle may be more difficult to endure. Google's competitors must educate the prospect, position their product, and, finally, convince the buyer that their "b" is better than the omnipresent Google.

Soople for Site Search

The Soople site allows a user to enter the URL of a specific public-facing Web site and run a query against that site. The result set shows only hits from that particular site. This is a narrowed Google search. However, when a Google Appliance indexes the same site, the result set for the same query will differ in many cases. The difference in the two result sets come about for several reasons. These include the depth to which the public Google spider and the Appliance's spider index, the frequency of the index refreshing, or the content "exposed" or "discovered" by each version of the Google search engine indexing the site. Do not underestimate the difficulty of explaining the differences. An organization can also use the public Google search engine to provide a site search service to its users, but using the public Google index will produce results that include Google advertisements. The Appliance does not. Note, however, that the Appliance can allow the licensee to display certain content in response to certain queries. These are "boosted" content and controlled by the licensee, not Google. In effect, the licensee can use Google's advertiser-support functionality to elevate content based on the licensee's perceptions of user interests and needs.

Lastly, Google is perceived in a very positive way. For example, Google is a good Internet citizen. Google started as an underdog in enterprise search, selling a search toaster when most vendors were selling boxes of parts for the customer to assemble. Google allows anyone to plug in an appliance and pay according to a simple price scale based on the number of documents indexed. When the appliance became available 2 years ago, it was labeled as simplistic. Now in version 2, the Google Appliance (see the "Google Appliance" sidebar on page 14) is more sophisticated yet still provides a quick, easy way to use Google search technology within an organization. The licensee can index content behind a firewall or create a custom search experience for the organization's public Web site.

Even its corporate offering seems competitively priced in the range of $40,000 up to $250,000. Many enterprise search systems weigh in with license and consulting fees in the million-dollar-plus range. Google seems to be an honest-to-goodness bargain. A Google competitor has an increasingly difficult time justifying an investment of orders of magnitude greater than Google. Even if the price were the same as Google's, the prospective enterprise search customer might still choose to give Google a whirl. As one government professional said, "Who is going to complain if we use Google to index our agency's documents?" About 20 years ago, the same comments were made about IBM.

Google's approach to enterprise search has weaknesses that go along with its ease of deployment, speed, and price scale. For example, search system vendors have successfully exploited such issues as these:

- How enterprise content differs from Internet content and why Google's supple metaphors don't apply in "enterprise content space." On the other hand, some search system marketers report that they have a challenge explaining "how and why" while keeping the prospect from "going into a coma."
- Google's "one size fits all" approach doesn't cover the diverse, highly particularized search needs of professionals in an enterprise. With only 15 percent of organizations operating an enterprise search system, the prospect may not know that research chemists and customer support representatives approach information in distinctive ways, and, therefore, each need a different search solution.
- Google's marketing touches lightly, if at all, on the issues associated with structured data residing in mainframe file systems or relational databases such as Oracle, DB2, and SQLServer. EasyAsk, Mondosoft, and Speed of Mind, among others, do a better job of handling content inside databases. Getting structured data from the organization's legacy system to an intranet index requires specialized processes and functions not supported by Google.
- The issues associated with indexing new and changing content so that the new information becomes available to users with acceptable query response time to the new and changed data. Google's approach implies that a query will be processed at "Google speed." Many organizations need specialized functions to keep data current and available. Plugging in another device sounds ideal, but the reality is licensees may need functionality, not boxes, to get what they want.

What's Google's response to this list of shortcomings? Wait for version 3.

What's Ahead?

The future of search occupies pundits from Ovum in the U.K. to the blue-chip providers of crystal ball gazing at Gartner, Jupiter, IDC, and dozens of other think tanks. A roundup of predictions makes interesting reading, and I will not identify the "prediction" with the firm because crystal ball gazers adapt quickly when events prove them wrong. Consider

these Nostradamus-like Promethean assertions:

1. Microsoft will triumph in search. Google will become tomorrow's Netscape. Microsoft has desktops, desire, and cash.

2. Search will disappear into other software. Search will become intelligent, and the results appear when and how the user requires the information. Endeca, InQuira, and other companies don't sell "search"; these companies sell solutions and bottom-line payoffs.

3. Consolidation will create a handful of winners who control lucrative accounts such as the Fortune 1000, the federal government, and well-heeled trade associations. The choices will be similar to buying an automobile from one of the major manufacturers.

4. Newcomers will "win." The amount of activity in Australia, Japan, India, China, and elsewhere sets the stage for the next revolution in the search ecosystem. Autonomy, Convera, Fast Search & Transfer, Verity, and now Google have to look over their shoulders. The next "search tsunami" may engulf the market any day now.

5. Search doesn't change. None of the Big Four have moved to a dominant position. Google and Microsoft will neutralize one another. The constant churn of newcomers provides a pragmatic way for the "big boys" to integrate new technology into their offerings.

6. Newcomers using social network technology similar to that found in Google's page rank algorithm or Eurekster's user-aware system or something equally innovative will offer a better way to index, classify, find, and display enterprise information.

None of these assertions is incorrect. None is 100 percent correct. Search is among the most complex tasks one can give to computers and software. Librarians and information professionals can still find jobs. Regardless of medium, most people — even those with college degrees — can't find information faultlessly every time they need it.

Engineers want to search by technical attributes, equations, and specialized terminology. Chemists want to search via structures like benzene or carbon. Lawyers want to search by a specific case attribute related to a general matter. Salespeople don't want to search at all. The system should display the prospect's options as the salesperson moves through the selling process.

One thing is certain, however: Google's entrance into the enterprise search market has begun the process of transforming the murky backwater of enterprise software into Half Moon Bay. Surf's up inside the enterprise. ◆