Envisioning the Future of Search

Stephen E. Arnold
ArnoldIT
Postal Box 320
Harrod's Creek, Kentucky 40027
sa@arnoldit.com
Voice: 502 228 1966

Drafted August 14, 2005. This essay forms the basis of Stephen E. Arnold's keynote at Information Today's search conference in New York City on September 28, 2005.

Introduction

After spending more than a year working on my monograph *The Google Legacy: How Google's Internet Search Is Transforming Application Software*, I have to resist the temptation to see the world through Google-colored glasses. (A sample chapter, a table of contents, and ordering information are located at http://www.infonortics.com/publications/google/google-legacy.html)

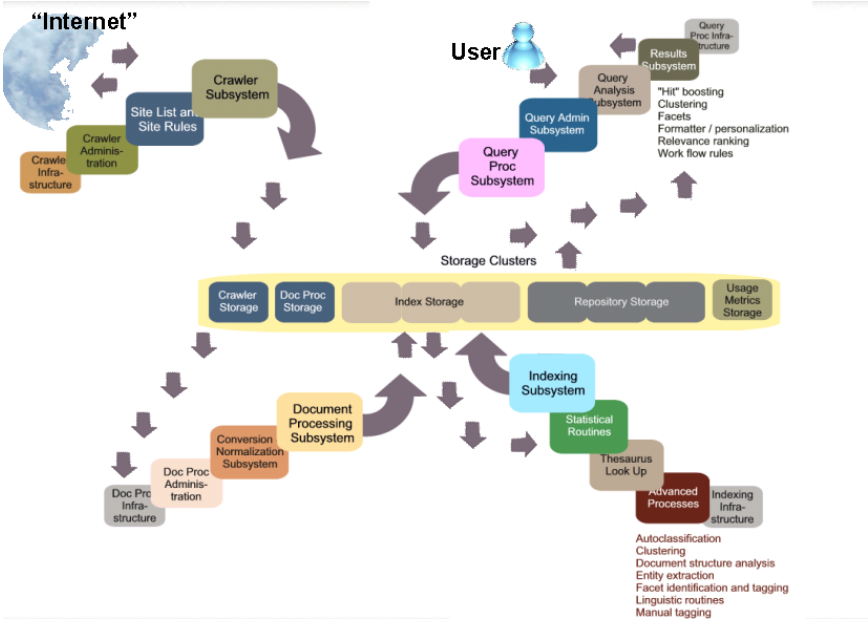
The future of search is by no means set in concrete. Jello is a more appropriate anchorage for the mind-boggling array of search systems, technologies, options, and approaches available today.

In the cheery past, search meant manual look ups in paper files, weighty reference books, and card catalogs that lead inevitably to printed documents.

Today, queries are handled by digital faeries. The results may be Web pages, Word files, music tracks, Power Point presentations, and Acrobat PDF documents.

Search is everywhere. It is part of the plumbing of the next-generation Windows operating system Vista. Oracle bundles text and data search tools with every Oracle database. Not to be outdone, IBM offers an industry-standard framework called unpoetically UEMA and the Omnifind "solution" that wraps in one package intelligent search, taxonomy generation, classification, and a batch of chocolate chip cookies. Google offers free Desktop Search Enterprise, the Google Appliance (a "search toaster"), and ad-supported "free" Web, image, news, and Usenet search.

Endeca, FAST Search & Transfer, and Verity have followed Hummingbird PCDocs and Open Text in creating enterprise applications that deliver search with a spoonful of financial sugar. The notion is that the bitterness of expensive remedies goes down easier when disguised as work flow or knowledge management.



The Hottest Search Trends

It will be useful to take a few moments and identify the search trends that are getting the most marketing hype and buzz. Keep in mind that every year or so, search, like a caterpillar sheds its cocoon and emerges with a different look and feel.

Much of what is “hot” or trendy in search is only slightly new. A bit of poking under the marketing promises, one finds string matching, thesauri, statistical relevance ranking, and algorithms that run more quickly on today’s fast, cheap hardware. That is not say there is nothing new in search. There are some surprising developments that warrant careful consideration. The trends that seem to have momentum include:

Ad-Supported Search

Users now expect advertisements in search results. The reasons for this are financial and practical. Search is expensive. Ads allow an organization to defray the cost of search and hopefully generate a profit. In the next six to nine months ads will appear in search results for colleges, associations, and some companies. The other reason is practical. Relevance of “organic” search results is generally poor. Even Google’s PageRank algorithm, now widely imitated by other search systems, flops when there aren’t enough votes to deliver a bang on result. Ads matched to a user’s query (in theory) provide another way to find relevant information in one click. Even work-flow search systems deliver preferred results, and these may be considered a type of advertising. After all, an Endeca system displaying the recommended mutual fund to a salesperson is little more than a financial product offering a higher payback to the brokerage firm using the Endeca system to enhance sales.

Federated Search

A federated search is a query that obtains results from two or more collections of information.[1] The term metasearch has fallen out of favor largely due to the confusion between metadata (data about information objects) and metasearch (a search across multiple indexes). Federated search takes the user's query and sends it to indexes for multiple collections of content or to different search engines. The query is passed against each of the "sources" and the results are returned to the user in a single, deduplicated list. The newest metasearch engines are remarkable beasts. Vivisimo, a spin out from Carnegie-Mellon University, federates, deduplicates, and returns results grouped in "folders" containing results related to a particular topic. Federated search is not a luxury. The number of search systems in an organization creates a need for a way to run one query across different collections of information.

Peer-to-Peer and Distributed Search

Peer-to-peer search or P2P as it is known in the popular press is an important innovation in search. P2P search freeware has contributed to the double digit decline in audio CD sales in the last two years. However, P2P search is extending its reach to other types of content, including video and text. The legality of file sharing is now before the U.S. Supreme Court. Regardless of the court's decision, peer-to-peer search is going to be important because distributed data technology and strong moves such as Microsoft's purchase of Groove Networks virtually ensures that peer-to-peer computing in some form will be will be a common feature of the enterprise search landscape.

Structured Data Search

Search is usually thought of as looking for articles and documents. In organizations, often the most useful information is held in structured databases. As mentioned elsewhere in this document, database software from Oracle, Microsoft, and market-leader IBM come with search tools. However, the queries must be crafted using Structured Query Language. Consequently, average users cannot search the complete contents of a legacy accounting system for past purchase orders from a Web browser. Extensible Markup Language, the "semantic Web", and clever software are unlocking information held in row-and-column format.

Stored Queries

Long a stalwart of Selective Dissemination of Information or SDIs, a new generation of search companies have used standing queries as the core of a search system. The idea is that "alerts" providing a user with new information on a particular topic has been extended to work processes. The idea is compelling. For each employee, figure out what search is needed to find the specific information required in a particular step in a work task. Program the query to fire automatically when the user reaches that particular step and display the results automatically. The payoff is eliminating the time and mental effort needed to formulate the query. Certain types of enterprise search can best be delivered using stored queries.

---

[1] The search system indexes one collection and creates an index. The second collection is indexed and a second index is created. A user wants to search both collections for information. A federated search system passes the query against the indexes and returns a single result list.

Text Mining

Like automatic classification and taxonomy tagging, text mining promises to the next “big thing” in search. The idea is compelling. Use the low-cost computing horsepower of today’s CPUs to analyze textual information. The text mining system will generate a report--usually in graphical or tabular formats--to reveal the main themes of the information, identify trends, and provide a bird’s-eye view of the important people, product, and events in the information. If this sounds too good to be true, it is, but there are valuable uses of text mining as a complement to traditional search systems. However, text mining is getting attention because it suggests that information does not have to be read by a human until an important nugget is identified.

Other Trends

No review of search would be complete without mentioning a number of other “trendlets” that could in the next year or so explode into full-blown trends. These include:

1. Automatic translation. This is more of a utility as opposed to a search system. The idea is that software can allow a user to query a corpus with documents in multiple languages and see the results in the user’s preferred language. Basis Technologies is a leader in tools for this type of search service.

2. Agent-based search. The idea is that software “learns” a user’s information needs and automatically formulates queries, obtains results, formats them, and then “puts them on hold” until the user needs the information. Vista includes tools to activate this type of search. Microsoft’s MyStuff is an example of this service.

Natural language processing. This is one of the Holy Grails of search. The notion is that a user can write a query as a sentence or paragraph (and in some developers’ implementation speak the query). The system than processes the query, runs the search, and returns the result as an “answer” or a spoken report. Rudimentary NLP is implemented in Ask Jeeves, Microsoft’s MSN Search, and Google now.

What’s Ahead?

For the next 12 to 24 months, nothing is more important than what I call the basics of search. There are defining one’s specific requirements, setting a budget (time and money), and selecting one or more systems that meet these requirements. The siren song of “free search” is melody to one’s ears. However, “free” systems are ultimately assigned a price tag.

The best plan of action for free and for-fee search is to:

1. Invest the time and effort to spell out what you need

2. Test different systems to gather data about those systems’ performance and suitableness to your situation

3. Select one or more systems but evaluate these systems’ performance on a regular basis

4. Repeat the process because search is a moving target.

Search, at least for the next 12 to 24 months, is a moving target. It is driven by elves--old and young--with different motives, technologies, and ideas about solving man’s oldest quest: finding the right information at the right time.