

Visualize This

BY JUDY LUTHER, MAUREEN KELLY, & DONALD BEAGLE -- 3/1/2005
FEATURES > INFOTECH FEATURE

Three students hunch over their computers, each doing research for a paper on nanotechnology. They all began their assignment with a simple Google search, only to learn that there are "2,500,000 English pages for nanotechnology." Their problem is not simply the number of pages available on the topic but also their very different interests: one is working on an assignment for biology class, one is studying engineering, and the third is studying business.

Google's success--and its challenge--is that it is an extremely efficient word finder; it builds indexes of the words on web pages. But words can be tricky things when it comes to conveying meaning. And a simple list of web pages containing those words doesn't provide us with much insight.

Visualization software is designed to help users get a "picture" of the meaning behind the words. Its underlying premise is that information retrieval benefits from smart organization of content. With visualization software, our three students could see their search results clustered to represent the different aspects of nanotechnology. They can then iteratively drill down in the appropriate cluster, all the while narrowing their search and learning more about their specific topic. As a result, those 2.5 million Google results become a gateway to discovery as well as to meaningful answers.

But will the new crop of visual interfaces displace the text-based tools to which we are so accustomed? Or will they be integrated with text-based search as just another option? Only time will tell if they provide enough of an advantage to sway users and don't end up just another new technology--interesting only to the media, scholars, and a few early adopters.

Awash in information

Information is now so accessible we are often thwarted by our own success in searching. When confronted with those thousands of search results on Google, we rely on its sophisticated ranking algorithms to bring the most important items to the top of the list. But if the answer is not on a "popular" web site, we may never see it on the fifth or 500th page of our results.

Good visual displays can compress information, convey context and relationships, and allow an array of options to be explored along alternate paths. It's the difference between reading a book to find the answer--with text organized according to perceived importance, of course--or looking at a table of contents and index. True, a well-ordered list can be very efficient. But the challenge of crafting an ordering schema that is right for all users in all instances has become increasingly difficult.

New visualization technologies may allow users to have greater personal control over the retrieval process. Just as someone in a physical library can view and draw clues from the arrangement and proximity of the books, visualization software can help users navigate through a virtual information collection.

Same search, differing results Information retrieval typically involves 1) selecting an appropriate resource, 2) searching it, and 3) examining the results to find suitable answers-- or, more often, documents or sites that contain those answers. At any point in the process, a user may decide to return to an earlier step to refine the results or try a different approach. Different types of searches place differing demands on the steps in this process. Some searches are focused, others are more exploratory. Sometimes the searcher wants everything on a topic: often he or she will settle for a "good enough" answer.

Visualization tools provide a compact, browsable overview of the search results in the form of topical clusters, graphs, maps, or other devices that convey themes by how they group the results. Instead of scanning a list of results sequentially based on their importance as ranked by the search engine, we can now see what topics are represented in our results and select one or more topical subsets to further explore.

"The human mind grasps visual representations more quickly than an equivalent amount of text," says Susan Feldman, IDC's research vice president of content technologies in the company's *Report on Interactive Data Visualization Tools*. "The eye seeks to compare similar things, to examine them from several angles, to shift perspective in order to view how the parts of a whole fit together."

Imagine you are searching for information on a type of wine, Shiraz, for example. This type of query is difficult to accomplish directly using a search statement because you may want to explore a range of countries--all of South America--or vintages or brands rather than a single one. Open-ended queries in search engines often produce a large results set, and the order may have little to do with the searcher's interests.

Beyond navigation of search results, some visualization tools provide a high-level topical overview of an entire information resource, such as HighWire Press's Topic Map, which displays more than 54,000 subjects and their hierarchical subheadings. It offers a way to bridge the search/browse dichotomy by supporting "browse-initiated querying" as well as "query-directed browsing." This type of guided browsing provides opportunities for the serendipity that is often lacking in search.

How they work

Visualization tools use two basic approaches to clustering information: they use metadata (such as cataloging information) that is associated with the information resource, and they use statistical and/or linguistic algorithms to create topical clusters on the fly. These approaches can be used alone or in combination to provide different views of the content.

They also differ in the type of visual metaphors that they employ. This can include minimal graphics based on nested lists or file folders; hyperbolic browsers; relationship diagrams using abstract shapes (circles, squares, lines) and connectors; geospatial maps, either abstract or actual; tables and graphs; time lines; or representations of real/concrete objects. For more on the different tools/companies, see "[The Visualizers](#)," p. 36-37.

Clever use of visual devices such as color, shape, size, position, and connection creates a multidimensional navigation space in which a lot of information can be conveyed in a compact display, including topics, relationships among topics, frequency of occurrence, importance, etc. While the elegance of the display can be appealing (or distracting), the real value comes from the users' ability to manipulate the display: to travel down a path, leaving breadcrumbs, collecting samples, and reversing and redirecting their steps. The excitement of a clever

display will wear off quickly if users can't control the exploration process. Success calls for flexibility in supporting different user styles and in extracting top value from different information collections.

One limitation with visualization software is whether users have adequate screen sizes to view the new, multicolumn displays properly. A recent trend is the shift by the young, mobile, Wi-Fi, connected generation away from desktop computers (and large screens) to laptops. John Sack, associate publisher and director of HighWire Press, observes that "it is important to take this reduced screen real estate into account when designing the interface for visualization tools."

Projects in development

The corporate sector has used visualization technologies as part of enterprisewide systems since the late 1990s, incorporating products from KartOO, TheBrain, and ThinkMap. But in libraries, the projects are just underway.

Groxis is partnering with both an academic and a corporate library to extend the applications of Grokker. SunLibrary Grokker enables employees to execute one search across IEEE, netLibrary and the web. Cindy Hill, SunLibrary manager, reports that engineers offered testimonials such as "simply amazing" and suggested enhancements--a true sign that they are vested in its development.

Stanford University is at work with Groxis on the development of Grokker E.D.U., which faculty, staff, and students can download and use to set up saved preferences. Michael Keller, university librarian, says that "the Grokker map is easy to navigate and allows users to quickly get to relevant results." It also "allows a form of federated searching" that "offers a way to bring together information from both the public and the proprietary environment in a single, comprehensive search with results displayed in a way that helps readers make sense of it all."

OCLC is implementing a data visualization pilot project in conjunction with Antarctica Systems Inc. to evaluate library users' experiences with search and display using a visual interface that offers the option to "try an alternative view of ebooks" in the Electronic Books database on OCLC FirstSearch. This will take users to a visual representation of the Electronic Books database, a static database of about 211,000 ebook titles. Included in this pilot are Mekko Maps, which depict all the subcategories of a taxonomy map within horizontal and vertical bars that show the user the subject categories they could choose.

Impact on libraries

Publishers such as the Institute of Physics and platform hosts such as HighWire have already incorporated visualization software to display their search results. Library users who find this approach useful may come to expect it as part of interfaces. It's also assumed that the generation known as "Millennials," now entering college, prefers visual over text-based systems. If popular consumer software, such as Google, adopt a visual interface, other large systems will quickly seem out-of-date.

In 2005, academic and public libraries in the United States will see several integrated systems offer visualization. Both TLC and VTLS are adopting AquaBrowser; Dynix is in discussions with them as well. Visualization tools may well raise the bar for all library systems vendors.

Visualization solutions may also increase the adoption of metasearch engines in libraries by

providing an appealing way to present search results from multiple databases. Michael Gorrell, CIO at EBSCO Publishing, confirmed that EBSCO is talking to Groxis to allow Grokker to work with EBSCOhost databases.

Kate Noerr at MuseGlobal notes that the growing presence of metasearch means that users are viewing lower-precision, higher-recall information, and the problem then becomes how to find the needle in the even bigger haystack. "Visualization tools, such as clustering, analysis, winnowing, refinement, etc., will aid significantly in reducing what a metasearch engine retrieves for users," says Noerr.

The path ahead

Visualization tools have received lots of positive attention from the popular press. When the new version of Grokker was released, David Kirkpatrick in *Fortune* wrote, "It makes me wonder if Google really does have search as sewed up as we often assume. When you use Grokker you realize just how brain-dead even the best search tools are today." And this January, CNN noted that these "intriguing technologies are getting better at bringing order to all that chaos and could revolutionize how people mine the Internet for information."

Academic librarians at many of the current test sites are cautiously optimistic. They report that users who try these tools generally find them useful for exploring search results, especially on topics at the periphery of their expertise. Users also value the opportunities for serendipitous discovery. However, these tools are not the first thing users turn to. "The key to success in implementing visualization tools is to understand your user," says HighWire's Sack. "Users are familiar with the Google interface. They will start with a keyword search no matter what you tell them to do. If the visualization tool is not a conspicuous, well-integrated part of the interface, it won't be used."

Today, it is ordinary users who determine the success of new technologies. The information market has become a mass market with revenues--both direct and indirect--that is dependent on a critical mass of users. The next several years will determine whether users will want to see it, as much--if not more-- than just read it.

Link List

The Associated Press. "Better Search Results Than Google?" CNN.Com, Jan. 5, 2005;
www.cnn.com/2004/TECH/internet/01/05/seeing.search1.ap

Beagle, Don. "Visualization of Metadata," *Information Technology & Libraries*, December 1999.

Beagle, Don. "Visualizing Keyword Distribution Across Multidisciplinary C-Space," D-Lib, June 2003;
www.dlib.org/dlib/june03/06contents.html

Kirkpatrick, David. "Going Deeper Than Google," CNN, December 17, 2003;
www.cnn.com/2003/TECH/ptech/12/17/fortune.ff.deeper.google/index.html

OCLC's E-books Pilot with Antarctica
<http://ebooks.antarctica.net>

Author Information

Judy Luther is President of Informed Strategies, a consulting firm; Maureen Kelly is a consultant and formerly Executive Director of Strategic Development for Nstein Technologies and VP for Strategic Development at BIOSIS; Donald (Don) Beagle is Library Director, Belmont Abbey College, NC

The Visualizers

Librarians and publishers are using and testing several software products. While they all share certain features, there are deep differences in philosophy, functionality, and market strategy among these young companies. Here are the major products, organized by type of visualization software. For more company information [see below](#).

Text clusters: Vivisimo

Founded in 2000 by research computer scientists at Carnegie Mellon University, Pittsburgh, Vivisimo's customers include corporations as well as a growing list of publishers such as Stanford University's HighWire Press, the Institute of Physics, American Association for the Advancement of Science, and British Medical Journal.

Vivisimo uses natural language rather than imbedded metadata to cluster search results on the fly. Icons are reminiscent of Windows, and results appear in traditional and familiar nested lists. Because the Vivisimo clusters are dynamic, they change as the underlying content changes. In a library context, Vivisimo would not offer support for clusters based on persistent classification schema like LCC and DDC.

Content Integrator is Vivisimo's federated search, or metasearch. When used together with the Clustering Engine, multisource search results can be filtered, merged, ranked, and presented in groups, or clusters, based upon content similarities.

Vivisimo's latest development, awkwardly named "Clusty," is in beta mode. It automatically groups search results into categories for the web, news, images, shopping, encyclopedia, and gossip from the tabloids and offers the option of customizing a display of results in selected categories.

Hyperbolic browsers: Inxight, xrefer

Inxight was founded in 1997 as a spinoff of Xerox Parc. With over 70 patents in natural-language processing and information retrieval, the company prides itself on managing unstructured data. Customers include Pfizer, IBM, Factiva, Thomson, and the LexisNexis Directory of Online Sources.

Hyperbolic browsers are used for topical browsing of large information collections. As users move down a branch of the categorization tree, the visual display shifts to reveal details while retaining a condensed view of the higher-level relationships. It relies on existing metadata to organize documents visually.

An example of a hyperbolic browser that librarians are likely to be familiar with was created by xrefer. Founded in 1999, xrefer specializes in the online delivery and metasearch of value-added reference information, combining texts from various publishers. The xrefer visual browser is a Java applet designed to provide browsable links that are context-relative and follow underlying semantic relationships.

Information maps: Antarctica Systems

Founded in 1999 by Tim Bray, Antarctica was inspired by Edward Tufte, author of *The Visual Display of Quantitative Information* (Graphics Pr., 2001). VisualNet is a server-side application that can be accessed by any user with a standard web browser. It uses cartographic techniques to map search results from existing hierarchies and their subsidiary clusters (such as LC classification). Because of this, VisualNet results tend to be very stable and the clustering predictable.

VisualNet can use a concrete metaphor but does not require it. Belmont Abbey College's Scholastica Project chose books on a shelf for its start screen metaphor, but once the user searches or browses beyond the start screen, the visualization becomes purely abstract. The project demonstrated a way to embed LCC captions (terms associated with each section of the LC classification scheme) as searchable elements of the hierarchical schema.

Graphical: Groxis

Groxis was founded in 2001, and its product, Grokker, was inspired by Robert Heinlein's term *to grok*, meaning to understand something at a very deep level. A multilevel application with server-side and client-side components, Grokker offers three levels of configuration that require the download of an applet.

Grokker displays colored circles within circles (or squares within squares), grouping classes and subclasses. Mouse-overs display the metadata accompanying each item such as print books, web sites, ebooks, and videos. The display goes through a period of shifting, during which circles appear, disappear, and rearrange at approximately two-second intervals as records are downloaded, until the batch processing is finished. Users can sort results by domain and customize and store results for future use. Flexible filters that adjust to the source being searched make it easy to narrow a search on the fly.

Hybrid visualization: Medialab Solutions

Medialab was founded in 1990 as the research facility of BSO Origin Philips and launched in Amsterdam in 2000 as Medialab Solutions BV. AquaBrowser Library is a search results navigation tool designed specifically for libraries and used by over 40 percent of the public libraries in the Netherlands and by the National Library of Singapore.

AquaBrowser Library incorporates a dictionary and thesaurus, which lets it recognize what is being searched for and offer relevant associated words, translations, and spelling variations and synonyms. AquaBrowser uses co-occurrence analysis to create associations that cluster the search results showing areas of interest using smart visualization. Users can refine their search using a nested list of natural-language terms plus categories from the MARC record.

Visualization Companies

A9.com, Inc.
Palo Alto, California.
A9.com, Inc.
Palo Alto, CA 94302-0504
www.a9.com

anacubis
6551 Loisdale Court
Suite 600
Springfield, VA 22150
Tel: 1-703-313-8875
www.anacubis.com

<p>Antarctica Systems Inc. 700 West Pender Street, Suite 1601 Vancouver, British Columbia Canada V6C 1G8 Tel: 604-873-6100 Toll Free: 1-866-638-6277 www.antarctica.net</p>	<p>Groxis, Inc. 30 Hotaling Place, Second Floor San Francisco, CA 94111 Tel: 1-415-398-0820 www.groker.com</p>	
<p>KartOO France www.kartoo.net Tel: 33 (0)4 73 28 98 26</p>	<p>Inxight Software, Inc. 500 Macara Avenue Sunnyvale, CA 94085 Tel: 1-408-738-6200 www.inxight.com</p>	
<p>Medialab Solutions BV Modemstraat 2B 1033 RW Amsterdam The Netherlands Tel: +31 (0)20 635 3190 email: info@medialab.nl www.medialab.nl/</p>	<p>Nstein Technologies 75 Queen Street, Suite 4400 Montreal, Quebec, Canada, H3C 2N6 Tel: 514-908-5406 Toll free: 1-877-nstein-1 www.nstein.com</p>	
<p>TheBrain Technologies Corp. 2001 Wilshire Blvd. Suite 310 Santa Monica, CA 90403 Tel: 1-310-656-8484 www.thebrain.com</p>	<p>Thinkmap, Inc. New York, NY Tel: 1-212-285-8600 www.thinkmap.com</p>	
<p>Vivisimo, Inc. 2435 Beechwood Blvd. Pittsburgh, PA 15217 Tel: 1-412-422-2499 www.vivisimo.com</p>	<p>xrefer Ltd 31 St James Avenue, Ste 370 Boston, MA 02116 Tel: 617 426 5202 Toll-free: 877 426 5202 www.xrefer.com</p>	