

Internet Insights

by Péter Jacsó

New Web Similarity Search Tools

Are these much-touted search engines panacea, snake oil, or in-between?

Web frenzy has been with us for years, but lately it has reached new heights. Hollywood must be watching with horror as Web-related news steals front page headlines and prime TV spots from Tinseltown gossips. Veteran anchormen of the evening news seem to compete to see who can utter more and longer URLs without pausing for a breath.

I would not be surprised to see someone start a rag called *Worldwide Enquirer*, and Jerry Springer put on a show with the topic "Web Abusers in the Family." Neophyte enthusiasm for everything that has the word "Web" in it is understandable, but it is alarming when people who should know better lose their common sense.

You're Charging a \$2 Penalty for What?

As I was working on this piece, I heard on the news that Delta Airlines announced that it is going to add a \$2 charge to every ticket not bought through the Web as a penalty for the non-Webheads. This news was so absurd that *U.S. News & World Report* first reported the opposite of it, and then had to run a correction in its February 1 issue that the surcharge applies to domestic round-trip tickets that are *not* purchased on its Web site. Delta must have been so excited about putting up a com-

scheduled flights. But I have been doing that for 20 years, honing my skills on the rather primitive OAG—Official Airline Guide—database instead of on the many nifty travel sites, not including the one from Delta.) Perhaps Delta should have offered a \$2 discount for those who use the Web site, as it saves the company processing costs, and that announcement would have had a much better ring to it.

The spell of the Web also works the other way. As a prelude to my new *Savvy Searching* column in the February issue of *Online & CD-ROM Review*, I discuss the fear-mongering articles written lately that suggest that anything that comes from the Web should be suspect. If you have been reading me in this decade, then you know that I say the same thing about many high-price-tag databases created and hosted by information industry stalwarts. Ironically, those who pen these articles and make these speeches mostly credit (if they credit at all) Web publications, ignoring the rich traditional print literature by long-time database-quality crusaders familiar to readers of *IT*, such as Reva Basch, Anne Mintz, Ruth Pagell, Carol Tenopir, Jeff Pemberton, Nancy Garman, or Barbara Quint—to name but a few. The unprecedented unanimity of Congress in passing the Communications Decency Act of 1996 was surpassed only by the unprecedented incompetence of those who tabled the bill, and the (perhaps) unprecedented ignorance in Internet matters of those who voted for it, which was almost the entire Congress. Luckily, the Supreme Court recognized the brutal unconstitutionality of the law and struck it down. Of course, there is a new round now with CDA 2 still pending.

What made me write this tirade is the news about some over-hyped Web search products in the printed press and on the Web. They certainly have merits and deserve interest, but not like the frenzy of teenage girls for the handsome Hansons. I take on Direct Hit this month, and I'll come back with news about others in later issues.

Not So Direct and Not So Much of a Hit

Direct Hit is the startup company that won the respected MIT \$50,000 Entrepreneurship Competition with its Popularity Engine. It is already available as a sidekick engine for mainstream search engines powered by Inktomi, such as HotBot or MSN, but can be also used in a stand-alone mode on <http://www.directhit.com>. The Popularity Engine sits in the background tracking searches, collecting data of queries, and ranking sites that users selected from the result lists of those queries. The queries are labeled by a descriptive name. When you make a search on the search engine that also hosts Direct Hit and the results are displayed, there are hotlinks on the top (to take you shopping). Clicking on the hotlink "Take the Top 10 Most Visited Sites" will show

the titles of searches that were deemed to be related to your search along with the list of the 10 sites selected most often by those who made the same query (see Figure 1).

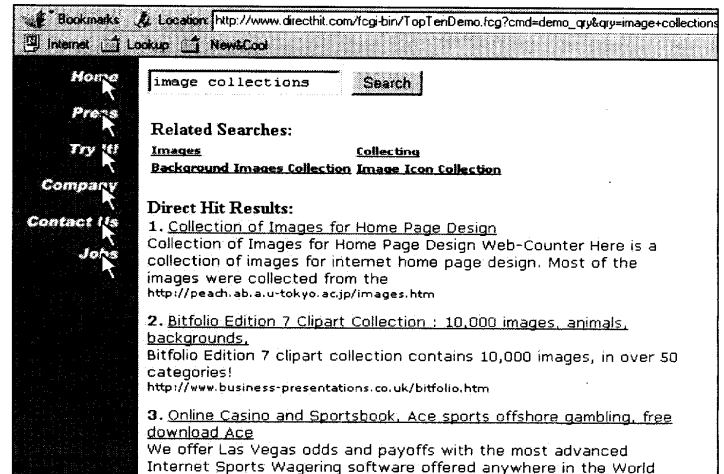


Figure 1

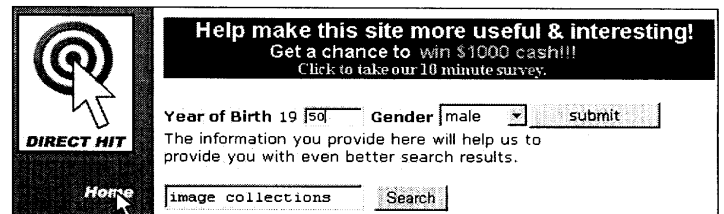


Figure 2

If your query does not match exactly any of the previous query titles, then only the related titles are listed. For example, there was no perfect match for my "Renaissance Art Collection" search, so there was no list, but most of the Related Searches were close enough. If there is a match, the Top 10 is often less than 10. For example, Popularity Engine yields only seven sites for the query "image collections" (and five more-or-less-related search titles—to be fair). This is not a problem if the seven sites are really relevant. However, I did not find these to be the most relevant sites out of those that I know, and the third-ranked site—about the most advanced sports-wagering software—did not instill in me that immediate "Eureka!" feeling.

The point is that the best quality, manually compiled Web directories yield better and more comprehensive results—although without ranking. Then again, the sunglass symbols in Yahoo's directory are better indicators for must-see sites than Rank #1 of Direct Hit. It is a good tool, but not the panacea it is heralded to be.

Image Collection Preferences of 49-Year-Old Males?

That still would not make me raise my eyebrows, but it certainly would make my

eyes browse the classified Web directories and the URL collections of real image experts like Paula Berinstein or Howard Besser rather than Direct Hit's Top 10 List. What makes me skeptical is the way Direct Hit tries to make the results more relevant: It encourages you to supply your year of birth and gender to customize the results (Figure 2).

Will that really help Direct Hit to provide even better search results as they claim? I doubt it. I'll leave it to your imagination

“
Similarity searches
by such criteria as
same author, same
subject headings, or
their combinations,
on the other hand,
are indeed far more
likely to yield
relevant results.”

mercial Web site that it believed that all Uncle Joes and Aunt Janes will immediately drop their phones and travel agents (who don't get even the much-reduced commission for tickets bought on the Web) and log on to <http://www.delta-air.com>.

It ain't gonna happen. I hope that customers will revolt and take their business elsewhere. (This is not envy-talk. If there is something I think I do really well, it is finding the cheapest airfares on the Web for

what image collections other 49-year-old males might be interested in. I don't want to sound pious like a televangelist, but I can't refrain from wondering what is common in terms of image collection preferences between me and those 49-year-old male Web users who pop open a six pack of beer while perusing the more notorious image collections the Web has to offer. I am not sure that it would be analogous to my searching for image collections on the Web to prepare for my next class or for my next column—while popping open my single can of beer.

Similis Simili Gaudet

This still would not concern me if I had not been reading everywhere lately that this is the next big thing in the search tool arsenal. Music sites, movie sites, and Web book stores have been offering such "similis simili gaudet," or "like takes pleasure in like," features for quite some time, and I never considered them important. I love the Amazon.com site, but not because it offers a button to see what other books were ordered by those customers who ordered the same book that I am looking at. It may be relevant, it may not be. The customers may have been shopping for their families

(continued on page 29)

The Gale Group

(continued from page 27)

we've blended three distinct approaches to reference materials into an integrated system that offers students the best of three worlds."

Entering through the School Collection home page, students are met with topic headings that direct their searches. Results from all databases are delivered at once, allowing students to travel throughout unique databases—gathering a variety of documents—without barriers.

Each collection features appropriate resources from InfoTrac (which delivers daily updated articles from as many as

350 newspapers and magazines), The Gale DISCovering Program (comprehensive, multi-source references correlated to national curriculum standards), and primary documents from American Journey: History in Your Hands. For example, a student's search on the Civil War would include recent newspaper accounts of modern perspectives on Civil War history, the text of the Emancipation Proclamation, letters from soldiers to their families, specially written articles on the Reconstruction Era, and biographies and photos of John Wilkes Booth and Abraham Lincoln.

Gale says that The Gale Group School Collections will be the first in a series of

services that integrate data from the former Gale Research, Information Access Company, and Primary Source Media.

Literature Resource Center Adds Literary Journals

The Literature Resource Center, Gale's online literary research service, now includes full-text articles from 24 important literary journals including *American Poetry Review*, *Review of Contemporary Fiction*, and *World Literature Today*. While the former IAC has been providing online access to these journals, this is the first time they've been integrated within a literature research service.

Launched in late 1998 (see the September 1998 issue of *IT*, p. 27), the Literature Resource Center has continued to grow and now includes along with the full-text journal articles 90,000 biographies, criticism on all generations of literary works, plot summaries and explications for more than 2,000 works, recommended reading lists for the most-studied authors, the full text of *Webster's Encyclopedia of Literature*, and 12,000 articles and critical essays on 2,500 of the most-studied authors in high school and undergraduate programs, according to the announcement.

Source: The Gale Group, Farmington Hills, MI, 800/877-GALE, 248/699-4253; <http://www.gale.com>.

Internet Insights

(continued from page 28)

and friends. The guy who bought the same book for himself that I am contemplating may have also bought another one or two for himself about organic foods that I could not care less about. Neither would he care about my book on the differences of inter-gender metacommunication in Europe and the U.S. that I picked up along with the book about Adobe Acrobat.

These similarity options, however, were not hyped breathlessly. Similarity searches by such criteria as same author, same subject headings, or their combinations, on the other hand, are indeed far more likely to yield relevant results. These were not pioneered by talented Young Turks of the Web but they improved what OCLC, Silver-Platter, Ovid, UMI, and DIALOG (on CD-ROM only) offered for a long time under the name of lateral searching or sideways searching since the 1980s.

Direct Hit can contribute to the betterment of Web searching, but statements I found on the *PC Magazine* Web site, such as "The Popularity Engine solves the problem of searching on a general topic and receiving thousands of links through which you have to wade to find what you want" would just make those guys lazy and be content selling snake oil like the age and gender optimizer. Direct Hit does not solve the problem of info glut, and such accolades do not prod it to go further.

Péter Jacsó is associate professor of library and information science at the department of information and computer sciences at the University of Hawaii. He won the 1998 Louis Shores/Oryx Press Award from ALA's Reference and User Services Association for his discerning database reviews. His e-mail address is jacso@hawaii.edu.

We welcome comments and suggestions from our readers.

Send your letters to the editor by mail, fax, or e-mail (hoffmand@infotoday.com)

or by way of our Web page located at <http://www.infotoday.com>.

Finding Methods...

Through the Madness.

Log onto methodsfinder.org and get to laboratory methods — fast! **MethodsFinder** is the only centralized source for methods information, including thousands of full-text protocols. The easy-to-use search engine directs you to the most current protocols from journals, books, researcher Web sites, commercial suppliers, and even meeting presentations.



MethodsFinder is the source for unique, hard-to-find methods information which is reviewed and expertly catalogued by BIOSIS' scientists for scientists. Browse and search by type of method, title, subject, disease, or organism. **MethodsFinder** keeps you pointed to new, better, and more economical ways to conduct research. So you can focus on your work.

FREE ONE-MONTH TRIAL.

Try **MethodsFinder** for a month by ordering one of four ways. Mention code **IT399MT**.

- e-mail to info@mail.biosis.org
- log onto www.methodsfinder.org
- call 1-800-523-4806, press 1 (USA and Canada), 215-587-4847 (Worldwide)
- fax your request to 215-587-2016

MethodsfinderSM

Biological lab methods at your fingertips.



www.methodsfinder.org