



## BUILDING BETTER SEARCH ENGINES

By Pam Frost Gorder

**F**ilippo Menczer remembers the day in 1993 when he downloaded his first Web browser. He was a graduate student studying artificial intelligence (AI) at the University of California, San Diego. The Web was still relatively new, and there were no search engines to sort through it; users could only follow links from one page to another. Finding a specific piece of information required a combination of determination and serendipity. As he clicked through his first series of links, Menczer thought, “we’re like ants searching for food!”

Today, he’s an associate professor of informatics at Indiana University, and he studies how the Web grows and evolves. Search engines have profoundly changed the way we use the Web, he says. Finding information is easy—so easy, in fact, that one common notion about search engines is that they bias Web traffic by directing people to popular sites and away from less popular, yet relevant, sites.

When Menczer set out to study search engine bias, he didn’t doubt that it existed; he just wanted to see how it worked. To his surprise, he and his colleagues found very little evidence of it. This research, which they published in the *Proceedings of the National Academy of Sciences* (vol. 103, no. 34, 2006, pp. 12684–12689), hints at the challenges that scientists face as they develop new and better ways to help us find information.

### A Living Web

Menczer’s diverse academic background—degrees in physics, cognitive science, and computer science—isn’t unusual among his colleagues in informatics who are interested in modeling complex systems. The Web is just such a system, with many components that come together with no centralized control, yet with an emerging *scale-free* structure. Communications networks, social networks, and even protein–protein interaction networks in biology are all scale-free networks. Although there’s no central control, patterns do emerge, in that connections between networks nodes tend to favor certain popular nodes, or *hubs*. In this way, the Web is like a natural, growing system. Menczer and his research team used what they knew about scale-free networks to model how search engines affect the Web.

They examined a particular search engine algorithm called PageRank, one of the methods that Google uses to rank its search results. The more people visit a Web site and link to it, the higher it ranks in PageRank. The researchers compared two model scenarios—one based on random queries to the PageRank algorithm, and another in which people simply browsed the Web using random links—with data from real-life searches performed on another search engine, AltaVista. Both situations showed signs of bias, with browsers tending to favor the more popular sites to the exclusion of less popular ones. But when the researchers looked at the real-life queries, the data showed much less bias.

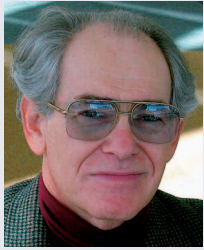
That’s because in real life, people don’t browse the Web in a random way, Menczer says. As long as users know what they’re looking for, and they type a specific topic into a search engine, they’ll be able to find sites—even new or less popular ones—that cater to their interest. From that perspective, search engines enable a kind of “survival of the fittest” Web paradigm. When relevance is important, even less popular sites rise to the top of the results list. And as more people visit them, these sites can become popular, too.

Although he got an unexpected result, Menczer says this study could still pave the way for better search engines in the future. “We originally thought that if we found a bias effect, then the search engines could modify their algorithms to compensate for it,” he says. “We didn’t find that bias, but the better you understand a network environment, the better you can model it and develop tools to better manage it.”

### Searching for Science

One person who isn’t surprised by Menczer’s results is Peter Norvig, Google’s director of research. The study “reflects what our intuition would have been,” he says. Whenever people search across broad categories, some very popular Web sites will emerge. But in practice, he adds, “people’s information needs are so varied that we’re sending them everywhere.” Still, he says, Google wants to do a better job of letting people describe what they’re

## OBSERVATOIRE LANDAU



### Petascale Simulations May Shed Light on Frailty of Human Condition

By Rubin Landau, Department Editor

In April, the US National Science Foundation (NSF) announced a solicitation for Accelerating Discovery in Science and Engineering through Petascale Simulations and Analysis (PetaApps; NSF 07-559). This competition will fund between 11 and 22 grants of up to US\$2 million each. It follows and supports a solicitation for creating a petascale computing environment for science and engineering (NSF 06-573), which is expected to award \$200 million in a single grant! Although the deadline for this latter competition has passed (missed your opportunity again?), and the award has yet to be announced, the competition for the \$2 million-and-less grants is still open.

In case you need reminding, deca = 10, hector =  $10^2$ , kilo =  $10^3$ , mega =  $10^6$ , giga =  $10^9$ , tera =  $10^{12}$ , peta =  $10^{15}$ , and exa =  $10^{18}$ . It's hard for me to imagine a computer capable of delivering sustained performance of greater than  $10^{15}$  floating-point operations per second on realistic and useful applications involving petabytes of data. Yes, we've witnessed Moore's law on our desktop computers, but at the current rate, it would take 30 years for my desktop to reach

the petascale, whereas the NSF proposal envisions such a system by 2011. To be accurate, the IBM BlueGene/L system has already achieved 281 Tflops, so it's not such a big jump to systems employing tens or hundreds of thousand of processors, each containing multiple cores capable of executing multiple threads and, often, arithmetic units that support small vector instructions.

Regardless of the technical achievements required to build a petascale computer, I suspect the reason for the latter NSF solicitation is that it's hard to imagine the type of problems that petascale computing can solve. Even if you could, the simulation, optimization, and analysis tools you would need to solve the problem simply aren't available. Scientists will have to devise algorithms that take advantage of the different types of parallelism available, with multilevel caches, local and main remote memory, intra- and internodal communication networks, parallel I/O, and the associated various levels of latency. We should be aided by the development of the new partitioned global address space compilers that offer simpler programming models such as co-Array Fortran, Unified Parallel C, and Titanium, together with their underlying native-mode communications library. But that's all part of our dreams for the future.

Clearly, you don't use such machines to compile your income taxes. Although we can always look at our old simulations with increasing resolution and accuracy, petascale computing will be most appropriate for new types of science and engineering problems that must be approached in new

searching for. Typing a few words into a small box amounts to an unnatural, one-way conversation with a search engine, and AI—a discipline in which Norvig has a good deal of expertise—could change that conversation. With AI, search engines could glean a deeper understanding of what people are looking for, and have a dialogue with them to help them find it.

Susan Dumais, principal researcher in the Adaptive Systems and Interaction Group at Microsoft Research, agrees that today's search engines give people limited opportunities to express what they want to find. She's using her own background in cognitive psychology to design programs that use people's natural memory skills to help them find things. Starting with the Windows Desktop Search program, she says, Microsoft is making it easier for people to retrieve useful items that they reference frequently. Now users can download Phlat (<http://research.microsoft.com/adapt/phlat/>), a free shell program that works with Desktop Search, to search everything on a hard drive—documents, email, pictures, music, whatever—not just by content but also by rich metadata, including user-generated tags. And the human brain is a great source of metadata.

“When you've seen something before, you have very rich memories about it,” Dumais says. “If you read an article, you may remember where it was published, what the article looked like, what else you were doing at the time, what people were involved in it. All of that information provides useful retrieval cues.” Menczer's study, she points out, showed that searching and browsing are two distinct but complementary ways to access information. Phlat tightly couples both activities: users can search using keywords or metadata, browse through categories of results, and refine their search based on whatever criteria matter to them. One day, people might use a product like Phlat to browse the Web. But there are other improvements Dumais would like to see in search engines, too—ones that would particularly benefit scientists.

“As a researcher, I gather information, but that's only the first step,” she says. “I analyze it, contrast it with other things, and then use it in writing a paper or making a presentation or sharing it with colleagues. We are doing a reasonable job of helping people gather information, but I don't think we're doing as good a job at helping people analyze that information. That's how I think we can really help the science and engineering community.”

ways. As a class, these new problems would involve multi-time and space scales, and multiphysics. They include, but aren't limited to, topics such as

- the radiative, dynamic, and nuclear physics of stars and the collision of stars;
- reactions of large biomolecules assemblages, such as cell membranes;
- nonlinear interactions between cloud systems, weather systems, and the climate;
- prediction of 3D protein structures from their primary amino acid sequence;
- determination of the Earth's internal structure via seismic inversions;
- design of molecular electronic devices;
- generation and evolution of magnetic fields in planets and stars;
- galaxy formation and evolution;
- design of specific catalysts, pharmaceuticals, and molecular materials; and
- climate modeling.

No doubt you're asking yourself, "where in this list is the human frailty that was advertised in this column's title?" I think it's us and our upbringing. If pushed, I might be able to imagine petascale simulations, although I suspect that truly imaginative petascale applications will need to come from the next generation of scientists and engineers who

were raised to think about problems on this scale. This leads me to the second half of this column. How do we prepare the human resources needed to provide the creativity and originality in petascale computing? The immediate answer is with educational activities focused on trends in high-performance computing. Three such educational programs are

- 2007 US Department of Energy Summer School in Multi-scale Mathematics and High Performance Computing (<http://multiscale.emsl.pnl.gov/>), 29 June–3 July 2007, Oregon State University. The summer program provides introductions to mathematical and computational methods used to model physical systems at various scales, tutorials, instructor-led lab activities, and research talks.
- SC07 Education Program ([www.computationalscience.org/workshops/summer07/index.html](http://www.computationalscience.org/workshops/summer07/index.html)), 10–13 November 2007, Reno, Nevada. Hands-on activities include how to apply computational science, grid computing, and high-performance computing resources in education.
- TeraGrid '07 ([www.union.wisc.edu/teragrid07](http://www.union.wisc.edu/teragrid07)), 4–8 June 2007, Madison, Wisconsin. The conference's theme is "Broadening Participation in TeraGrid." It features scientific results from the use of TeraGrid and tutorials on TeraGrid resources, such as visualization tools, Science Gateways, and Globus middleware.

See ya there!

## Value Chain

Both Norvig and Dumais say that researchers face unique obstacles when they search for information—the things they want to find, such as journal articles or research data, often aren't available for free on the Web. Christine Borgman, professor and Presidential Chair in Information Studies at the University of California, Los Angeles, knows why those items aren't available. She also adds that researchers have more control over the situation than they know.

Most copyright agreements allow researchers to post their articles online, albeit with some restrictions, but most researchers just don't do it—they don't realize they have the right, or they don't want to bother. Either way, they're limiting the number of people who can see and reference their work. And researchers tend not to post their data online because they're afraid they'll lose control over it, and people will use that data without crediting its source. If tools were created to make depositing data on the Web easy, while tagging the data to clearly indicate its origins, the situation might change. With open access to journal articles—even drafts and preprints—as well as data, researchers could assemble the "value chain" for a particular line of research.

Still, just getting at the scientific information that's already out there isn't easy. Borgman says this is because search engines are oriented toward the naïve searcher, not the scientific searcher. "As long as people are relying on search engines that are getting their money from finding the cheapest deal on airfares and cameras, they're never going to be very useful for science." To be really useful, search engines would have to reveal the existence of data repositories so that researchers could then perform specialized searches inside them. Libraries around the world are expanding their digital archives right now, she points out. If search engines don't reveal these archives, many treasures could go undiscovered.

## A New Search Engine

At Indiana University, Menczer is creating a new kind of search engine that's based on a very familiar form of Web metadata—bookmarks. GiveALink ([www.givealink.org](http://www.givealink.org)) is similar to the del.icio.us (<http://del.icio.us>) social-networking site in that it lets registered users share their bookmarks online and tag them with even more meta information. It's also like the news aggregator Digg ([www.digg.com](http://www.digg.com)), in that it lets users vote for the Web sites they think are important. Visi-

## DEPARTMENT EDITORS

**Books:** Mario Belloni, Davidson College, mabelloni@davidson.edu  
**Computing Prescriptions:** Isabel Beichl, Nat'l Inst. of Standards and Tech., isabel.beichl@nist.gov  
**Computer Simulations:** Muhammad Sahimi, University of Southern California, moe@iran.usc.edu, and Dietrich Stauffer, Univ. of Köln, stauffer@thp.uni-koeln.de  
**Education:** Michael Dennin, Univ. of Calif., Irvine, mdennin@uci.edu, and Steven F. Barrett, Univ. of Wyoming, steveb@uwyo.edu  
**News:** Rubin Landau, Oregon State Univ., rubin@physics.oregonstate.edu  
**Scientific Programming:** Konstantin Läufer, Loyola University, Chicago, klauffer@cs.luc.edu, and George K. Thiruvathukal, Loyola University, Chicago, gkt@cs.luc.edu  
**Technology:** Michael Gray, American University, gray@american.edu, and James D. Myers, Collaborative Systems, NCSA, jimmyers@ncsa.uiuc.edu  
**Visualization Corner:** Claudio T. Silva, University of Utah, csilva@cs.utah.edu, and Joel E. Tohline, Louisiana State University, tohline@rouge.phys.lsu.edu

## STAFF

**Senior Editor:** Jenny Stout, jstout@computer.org  
**Group Managing Editor:** Steve Woods  
**Staff Editors:** Kathy Clark-Fisher, Rebecca L. Deuel, and Brandi Ortega  
**Contributing Editor:** Cheryl Baltes and Joan Taylor  
**Production Editor:** Monette Velasco  
**Publications Coordinator:** Hazel Kosky, cise@computer.org  
**Technical Illustrator:** Alex Torres  
**Publisher:** Angela Burgess, aburgess@computer.org  
**Associate Publisher:** Dick Price  
**Advertising Coordinator:** Marian Anderson  
**Marketing Manager:** Georgann Carter  
**Business Development Manager:** Sandra Brown

## AIP STAFF

**Circulation Director:** Jeff Bebee, jbebee@aip.org  
**Editorial Liaison:** Charles Day, cday@aip.org

## IEEE ANTENNAS AND PROPAGATION SOCIETY LIAISON

Don Wilton, Univ. of Houston, wilton@uh.edu

## IEEE SIGNAL PROCESSING SOCIETY LIAISON

Elias S. Manolakos, Northeastern Univ., elias@neu.edu

## CS PUBLICATIONS BOARD

Jon Rokne (chair), Mike Blaha, Angela Burgess, Doris Carver, Mark Christensen, David Ebert, Frank Ferrante, Phil Laplante, Dick Price, Don Shafer, Linda Shafer, Steve Tanimoto, Wenping Wang

## CS MAGAZINE OPERATIONS COMMITTEE

Robert E. Filman (chair), David Albonesi, Jean Bacon, Arnold (Jay) Bragg, Carl Chang, Kwang-Ting (Tim) Cheng, Norman Chonacky, Fred Douglas, Hakan Erdogmus, David A. Grier, James Hendler, Carl E. Landwehr, Sethuraman (Panch) Panchanathan, Maureen Stone, Roy Want

## EDITORIAL OFFICE

COMPUTING IN SCIENCE & ENGINEERING  
 10662 Los Vaqueros Circle, Los Alamitos, CA 90720 USA  
 phone +1 714 821 8380; fax +1 714 821 4010; www.computer.org/cise/



## WEB TRENDS

For a brief look at current events, including program announcements and news items related to science and engineering, check out the following Web sites:

- 2007 Summer School on Computational Materials Science: Quantum Monte Carlo (QMC) from Minerals and Materials to Molecules ([www.mcc.uiuc.edu/summer-school/2007/qmc](http://www.mcc.uiuc.edu/summer-school/2007/qmc)). This summer program brings together scientists from the fields of geophysics, physics, chemistry, and materials science to learn about QMC calculations and their applications. The program will run 9–19 July 2007 at the University of Illinois at Urbana-Champaign.
- Accelerating Discovery in Science and Engineering through Petascale Simulations and Analysis (PetaApps; [www.nsf.gov/publications/pub\\_summ.jsp?ods\\_key=nsf07559](http://www.nsf.gov/publications/pub_summ.jsp?ods_key=nsf07559)). PetaApps is accepting proposals that develop petascale simulations and analysis tools. Researchers must demonstrate their proposals require petascale computing. The proposal deadline is 23 July 2007.
- Broadening Participation in Computing (BPC; [www.nsf.gov/publications/pub\\_summ.jsp?ods\\_key=nsf07548](http://www.nsf.gov/publications/pub_summ.jsp?ods_key=nsf07548)). This program seeks to increase the number of students pursuing post-secondary degrees in the computation fields, with an emphasis on increasing the participation of women and minorities. Deadline for proposals is 4 June 2007.
- CreativeIT ([www.nsf.gov/publications/pub\\_summ.jsp?ods\\_key=nsf07562](http://www.nsf.gov/publications/pub_summ.jsp?ods_key=nsf07562)). The CreativeIT program is seeking proposals that explore the study of creativity to further computer science and create new models of computational processes and approaches to education. The deadline for proposals is 21 September 2007.
- Fernbach Award (<http://sc07.supercomputing.org/?pg=awards.html>). Nominations for the IEEE Computer Society's Sidney Fernbach Memorial Award can be submitted at [www.computer.org/awards](http://www.computer.org/awards). Deadline for submissions is 30 June 2007. The award will be presented at SC07 in November.

tors upload their bookmarks, and then Menczer's team applies a machine-learning algorithm that maps out relationships between the bookmarks and creates a ranking scheme for the search engine.

For instance, if many people have the same two Web sites stored in the same folder in their bookmarks file, then GiveALink ranks the two sites as closely related. The aggregate data forms a semantic network, or ontology, for the Web. "It's a way to build a classification automatically, by everybody just submitting a piece of the puzzle,"

Menczer says. "So now I get a more trusted notion of a link, where there is meaning attached to it." Links are weighted to take into account people's notions of how things are related. This means that if any group of people, such as scientists in a particular discipline, donated all their bookmarks, they would build an ontology that was unique to them.

In return for donating their bookmarks to science, users can get personalized recommendations for everything from music to news to podcasts. And Menczer thinks industrial partners could integrate GiveALink's methods into any search engine to help them see relationships among results. The algorithm would be especially useful for navigating very large databases. "Ultimately, this technique will lead to even better ranking measures, in my opinion," he says. So far, the site has received nearly 2,500 donations and uncovered almost 5 million semantic links between pages.

**M**enczer's research group is continuing its studies of Web evolution. One project looks at Wikipedia ([www.wikipedia.org](http://www.wikipedia.org)), a Web encyclopedia constructed entirely by volunteers who create entries and edit them. Menczer wants to see how Wikipedia's network structure, in which many people can edit the same document, differs from that of the Web at large, in which most people have sole control over their own space. Yet another project has obvious commercial applications. Menczer's trying to determine what makes for a successful Web site—the look or the content, or more subtle factors, such as the frequency with which people link to it. It all comes down to human behavior.

"The more we understand what people are doing and how they're doing it, the better we can exploit that information to build better Web tools," he says.

CiSE

**Pam Frost Gorder** is a freelance science writer based in Columbus, Ohio.

When it comes  
to predicting,  
you can't afford  
to play around.



Making predictions can be tricky. Whether you're conducting research, detecting fraud, or diagnosing illness, drawing conclusions from limited data is serious business. NeuralTools from Palisade brings sophisticated Neural Networks to Microsoft Excel, so you can use your known data to make predictions with uncanny accuracy. Plus, NeuralTools Live Prediction updates results in real-time, making it the tool of choice among professionals. Anything else is just a toy.

**"Super Quick!"**

"I am delighted with the program for its speed and easy handling. The program is super quick."  
— Dr. Jose R. Iglesias-Rozas, Katharinenhospital

**"Awesome!"**

"Most folks are not engineers, but they can use NeuralTools to facilitate their own forecasts... just awesome!"  
— Barb Tawney, University of Virginia



**In North America:**  
1 800 432 7475  
+1 607 277 8000  
sales@palisade.com

**In Europe:**  
+44 1895 425050  
sales@palisade-europe.com

**In Asia-Pacific:**  
1 800 177 101  
+61 2 9929 9799  
sales@palisade.com.au

Download a FREE trial version of NeuralTools at <http://www.palisade.com?cise>

CiSE readers **SAVE 50%** off NeuralTools Pro today  
when you mention this ad!