From the Editor in Chief...

Define: Valedictory

Robert E. Filman • Google • filman@computer.org

Last column. My congratulations to Fred Douglass of IBM Research, who takes over as editor in chief starting next issue. Fred will be ably assisted by Doug Lea of SUNY Oswego, who is continuing as associate EIC, and Siobhán Clarke of Trinity College, Dublin, incoming AEIC. I would like to thank the magazine staff for all their help over the past four volumes — especially the magazine assistant, Hazel Kosky, and the magazine’s managing and lead editors, Steve Woods and Rebecca Deuel — and the AEICs and editorial board for their support and great work. All the mistakes have been my fault.

For the past few months, my head has been wrapped up in issues of Internet search. My apologies for missing columns. At least in the future, you won’t be surprised by my absence.

How Would You Find It?

Steven Wright has noted that the problem with having everything is where to put it. With the Internet, we’re close to having everything. The problem isn’t so much where to put it but how to find it.

Back before the Internet, if you wanted to know something, you asked your friends or consulted a librarian. Your friends had the virtue of a good understanding of your interests, tastes, and expertise, and expertise and social awareness of their own. Librarians relied on structured organizations of information, such as the Dewey decimal system and card catalogs, along with laboriously hand-built indices. Each would be willing to engage in a conversation to elicit what exactly you were really looking for and had a good sense of what most documents meant.

One of the first approaches to the problem of finding things on the Internet, was, in a certain sense, to imitate the librarian’s notion of catalogs. Early on, companies such as Yahoo built Web page directories. Want to find the computer science department at Stanford University? Start at the Yahoo “Education” link, wander down through the colleges and universities, then to US, California, and Stanford, and scan the results. This had the virtue that you would not only get to a page labeled “Stanford Computer Science Department,” but that you would get to the real page of the Stanford computer science department. Having reliable people create the directory made the overall results trustworthy. This scheme had the disadvantage that if you’ve got everything — a whole wide world of Web page makers — they’re likely to be busily creating pages faster than the reliable people can catalog them, even if you have bubble money to pay those reliable people.

The alternative was search engines. Search engines are programs (easily parallelizable programs at that) and can find and classify text much more quickly than (if not as accurately as) people. A search engine has two important parts: a crawler and an indexer. The crawler reads Web pages, finds which Web pages these pages reference, and then, recursively crawls the “promising” pages. The indexer builds some database representing these pages and their contents and, in response to a query, reports back the “best” pages for that query.

Crawlers face problems such as recognizing aliases for pages; dealing with dynamic, forbidden, and protected content; and scheduling their activities to both get their tasks done and not overwhelm any part of the Web.

Apologies to those devoted to crawling, but the most difficult issues are in indexing — efficiently organizing the information found by the crawl and formulating responses to queries.

Traditional information-retrieval (IR) mechanisms drove the first generation of Web indexers. IR is based on the premise that if a page mentions a particular word frequently, then that page is likely to be a good response to queries about that word. References in titles or boldface might be more valuable than casual references in the body of the text. For queries combining multiple words, traditional IR suggests paying more attention to the less fre-
quent words. That is, when looking for “Robert Filman,” prefer pages that have a lot to do with “Filman,” not those that do a good job on “Robert.”

In general, indexing systems have a certain amount of screen real estate for presenting their results; engines that present better results more prominently seem to produce happier users. Most simply, this can translate into just listing the results in order of perceived quality, but more generally it’s exhibited in areas of the display devoted to particular aspects of the results, tree-like link paths for some output, and so forth.

At dinner a few months ago, I spoke with a columnist for another publication who assured me that this Internet search stuff was just old-fashioned IR. He was wrong. Internet search is a full-contact sport, as similar to pure IR as football is to ballet. IR mechanisms work well in benevolent environments — when you can trust the documents in the corpus to actually be about the words they mention. This is fine for an internal database of documentation, but when the World Wide Web evolved into the world-wide market, having people look at your page acquired economic value. It had value even if it wasn’t the best page on the topic of their search, or, for that matter, even if it didn’t have anything at all to do with their search. The same economics that supports sending a million emails to make 60 sales supports having a million page views that produce 60 sales. So, littering a document that offers fake Viagra with references to Britney Spears, especially in yellow two-point type under a jpeg, can be profitable for the spammer, though aggravating for the 999,940 people who wanted to read about Ms. Spears, not drugs.

**Second-Generation Search Systems**

Directory systems produced good results because someone reliable vouched for their validity, but the number of Web pages and the variety of queries grew too quickly for manual methods to keep pace. This is especially true for rare and unpopular queries. Although IR techniques worked for a little while, they were too easily led astray by pages that pretexed about their contents. Even with nonspam pages, IR metrics aren’t a particularly good ranking method — frequency of mention is only loosely correlated with quality. The cleverness of second-generation search systems, such as Jon Kleinberg’s HITS and Sergey Brin and Lawrence Page’s PageRank, was to find a way to let a social consensus of recommendations influence the results. Most superficially, if a page has a link labeled “Robert Filman” with a pointer to http://home.comcast.net/~refilman/, then this page can be understood as a recommendation of http://home.comcast.net/~refilman/ as a good response to the query “Robert Filman.” Furthermore, the recommendations of “important pages” (such as the pages pointed to by many other pages) are more influential than the recommendations of unimportant pages.

Kleinberg took his ideas off to Cornell and was rewarded with a professorship and a MacArthur fellowship. Brin and Page used their ideas to start a company and were rewarded with places on the Fortune 20 list. If you want a far less superficial description of the mathematics underlying page rank and propagating importance, and a bit of gossip on the history of search engines, try Amy Langville and Carl Meyer’s book — you’ll come to understand that Internet search has blessed linear algebraists with a kind of social purpose that Rubik’s cube gave group theory and RSA, number theory.

Of course, if you get credit for having many pages point to your page with the words “Britney Spears” associated with the hyperlink anchor, then people who want to sell Viagra to Britney seekers will set up networks of mutually referent pages, each recommending the other for something for which they’re not organically appro-
The Future of Search Systems

What is the future of search systems? Your original system of librarians and friends understood what they were reading, could talk to you about what you were looking for, and had some sense of the social gestalt. Similarly, I can think of three broad dimensions for the next generation of search systems:

- The system could have a better understanding of what it finds. For example, a system that could not only recognize words in a Web page but also “understand” the meanings of those words could be quite useful. Unfortunately, the complete understanding problem is AI-hard. On the other hand, some qualities of partial understanding might prove useful.
- The system could have a better understanding of the user’s intent. This might come about either through a more thorough and explicit conversation about what the user is looking for (both in general and in particular) or, more subtly, by closer observation of the user’s behavior.
- The system could do a better job of applying the “wisdom of others,” for example, by understanding pages’ structural relationships, taking account of explicit recommendations, recognizing clues in user behaviors, and gauging the trustworthiness of such signals. Think of HITS and PageRank as first steps in this direction, and emergence of sites such as digg (www.digg.com, user-recommended news), stumbleupon (www.stumbleupon.com/, shared recommendations about Web sites), and del.icio.us (http://del.icio.us, social bookmarking), as well as techniques such as photo-labeling games (such as Luis von Ahn’s ESP Game, www.espgame.org, in which players try to find matching labels for pictures) as early steps in this direction.

One thing that will be true is that successful search engines won’t be the province of the faint-hearted or poorly funded. Running a popular search engine that produces the quick responses that users seem to like is a cap-
Smaller efforts can be interesting, but don’t scale. Consider Minerva, a peer-to-peer (P2P) Web search engine, in which peer computers keep their own page ratings, and the system synthesizes these peer opinions into an overall search. It’s a lovely academic project, which, unfortunately, wouldn’t work as a general, popular search engine. P2P systems are best at flouting regulations (for instance, music sharing in the face of copyright law), not supporting the trust relationships necessary for large-scale, pluralistic search. Spammers would have a delightful time with P2P search. It is true that such ideas might work perfectly well for small or closed domains. If you can restrict the network to trusted peers, or the network has too little traffic to be worth attacking, then P2P mechanisms could conceivably be viable. (After all, anything small enough can be too tiny to interest spammers — the Macintosh has less malware not only for being somewhat more immune to hacking than Windows, but also for having insufficient market-share to interest hackers.) Such small systems might be able to effectively combine individual wisdom on specific topics. However, if the goal is to produce a general and popular search engine, then capital and trustworthiness demand centralized control.

I n addition to the notes of appreciation in my introduction, I’d like to thank you, my readers (or at least the ones who have gotten this far) for your patience in putting up with my ramblings over the last four years.

References