The Scientist
Volume 19 | Issue 20 | Page 24 | Oct. 24, 2005 FEATURE | TRIBUTE

The Future of Citation Analysis

The challenge is to track a work's impact when published in nontraditional forms

By Jeffrey M. Perkel

THE HOUSE THAT GENE BUILT:

ISI headquarters, in Philadelphia. The building's facade is supposed to evoke moving punch cards, which the company originally used to store citation data. The following page lists the top 10 cited papers from the past two years, 10 years, and of all time.

In the 50 years since Eugene Garfield first proposed it,[1] the *Science Citation Index* has grown dramatically in size and influence. The database has expanded from 1.4 million citations in 1964 to 550 million today. Its list of source journals has grown from 613 to 15,721. And it has become a key tool for tenure, funding, and award committees.

The move to a Web interface that can analyze a century's worth of literature at the click of a mouse has made the *Science Citation Index*, now part of Thomson Scientific's *Web of Science* (*WOS*), more useful than ever. But the same Web that has given the *WOS* greater and greater power has also spawned publication avenues that leave open the question of how citation analysis will evolve in the near- and long-term.

Articles can be posted in multiple forms in multiple places: on the ArXiv.org preprint server, on the author's personal home page, and on the journal Web site, for instance. Those articles can be published almost immediately, giving the larger scientific community time to digest, incorporate, and ultimately cite them.

The *WOS* is more than a literature database; it measures how often journal articles are cited by others. How do you analyze all these new types of citations? "If you're trying to figure out the impact of that article, you've got to figure out how many links go to each source and bring them together," says Michael Koenig of the Palmer School of Library and Information Science at Long Island University, Brookville, NY.

Last autumn's launch of Google Scholar (GS) presents one solution. The free service searches and tracks citations to peer-reviewed literature (as the Web of Science does) and also conference proceedings, dissertations, pre- and postprint servers, and other nontraditional media. Last month, Yale University librarians Kathleen Bauer and Nisa Bakkalbasi published an analysis showing that GS yielded 4.5 more citations per paper on average in one journal, the Journal of the American Society for Information Science and Technology, for papers published in 2000, than did WOS.[2]

"We found that through Google Scholar, you do get a higher average number of citing articles, than you do through the Web of Science and Scopus. We were quantifying what you would have guessed," says Bauer.

But the scholarly value of nontraditional sources picked up by GS and not by WOS is yet unproven. And some major publishers, including Elsevier and the American Chemical Society, have declined to open their archives to GS, limiting its completeness. Peter Jacso, a professor of computer and information science at the University of Hawaii, Manoa, estimates GS has about 10 million source records to WOS's 35 million.

Jacso recently completed an analysis of GS, WOS, and Scopus, which suggests GS does "a really horrible job" matching cited and citing references, he says. GS "often can't tell apart a page number from a publication year, part of the title of a book from the name of a journal, and dumps at you absurd data."

For their part Bauer and Bakkalbasi write in their study that "ad hoc searches" in WOS, GS, and Scopus suggest their findings extend to other journals and other fields. They therefore advise researchers to consult GS in addition to WOS or Scopus, "especially for a relatively recent article, author or subject area." But they, like Jacso, note that until GS reveals precisely what it indexes and how often it updates, "it cannot be considered a true scholarly resource in the sense that Web of Science and Scopus are. An understanding of the material being covered is central to the validity of any search of scholarly material."

UNPUBLISHED BUT CRITICAL

Scientists can influence their peers beyond the published word, of course. Consider a scientist who develops a useful program, and posts it to a Web site from which it can be downloaded. Such contributions, says Blaise Cronin, the Rudy Professor of Information Science at Indiana University, are "subterranean, subcutaneous," and they are generally ignored in traditional citation analyses.

One place where they do sometimes appear, however, is in a paper's

acknowledgments. "By analyzing acknowledgements, you can demonstrate just how much people rely on one another, even competitors, and especially in the life sciences, where you are required to share reagents after publication," says Cronin, who has spent 15 years mining acknowledgements in scientific literature for their citation value. His recently completed analysis of acknowledgements in four years of Cell issues found that "over the course of three decades, the intensity of acknowledgment behavior rose for each category, most notably in the cases of materials (from 17.6% to 65.1%) and conceptual contributions (30.1% to 84%)."

To date Cronin's analyses have been painstaking, manual processes. But he won't have to work manually for long: This past December Pennsylvania State University researchers C. Lee Giles and Isaac G. Councill reported a systematic effort to extract and parse acknowledgement text from 335,000 computer science papers.[3] "Our work supports prior studies showing that acknowledgment trends for individuals do not correlate well with citation trends, perhaps indicating a need to reward highly acknowledged researchers with the deserved recognition of significant intellectual debt," the authors write.

Another metric that citation analysts are currently debating is the value of Web linkages (a link from one person's home page to another). Simply counting links isn't likely to be of much use, says Henry Small, chief scientist at Thomson Scientific and president of the International Society for Scientometrics and Informetrics. "Basically anything goes on the Web. You can have crackpots and charlatans linking to your stuff, [and] you can have Nobel Prize winners linking to your stuff."

Hypertext links reflect more informal, social contacts, says Small, while citations represent more formal expressions of intellectual debt. Nevertheless, he says ongoing efforts to map the Web, to visualize its connectivity and see who influences whom, are among the most sophisticated areas to evolve from traditional citation analysis. "People are attempting to use all the links to map the system of underlying communications or of ideas," he says.

Yale's Bauer suggests that with all the new options available, journals per se may

lose their dominance, in favor of the papers within them. The playing field could be leveled: Authors may not choose particular journals based on impact factors, but choose publishing methods based on effectiveness.

For now, however, the traditional refereed paper, wherever it happens to be published, remains the coin of the realm. Says Cronin: "As more of scientific literature moves to the Web and becomes available, you're going to have a richer picture of the life and vitality of a scientific paper than you can have today. So citation analysis won't become passé, it will become one of a battery of indicators with which to measure the impact and influence of a publication."

References

1. E Garfield "Citation indexes for science: A new dimension in documentation through association of ideas," *Science* 1955, 122: 108-11.
2. K Bauer, N Bakkalbasi "An examination of citation counts in a new scholarly communication environment," http://www.dlib.org/dlib/september05/bauer/09bauer.html*D-Lib Magazine* 2005.
3. CL Giles, IG Councill "Who gets acknowledged: Measuring scientific contributions through automatic acknowledgment indexing," *Proc Natl Acad Sci* 2004, 101: 17599-604.

The H-Index

Citation analysis data are routinely used to rank universities, departments, even countries. Yet the notion of using citation counts to rank individuals say, to determine whose grant gets funded is citation analysis' most controversial application. Recently, Jorge Hirsch, a professor of physics at the University of California, San Diego, described a new metric called the "h-index," which provides a sort of shorthand expression of an individual's scientific output and quality.[1] Hirsch explains that your h-index is equal to h if you have published h

papers, each of which has at least h citations. Papers with fewer citations don't count in the analysis. "So, it counts your significant papers," he says.

Hirsch calculated the h-index for several prominent life scientists. The top five were Sol Snyder (191); David Baltimore (160); Robert Gallo (154); Pierre Chambon (153); and Bert Vogelstein (151). Thus, Johns Hopkins University pharmacologist Snyder has 191 papers that were each cited at least 191 times. (Hirsch's h-index is 49.)

Though he doesn't endorse its use in awarding prizes, Hirsch does advocate using the h-index, in conjunction with other metrics, to award grants. "I think that is useful information that should be taken into account," he says," to know the resources that are being allocated are being well distributed."

References

1. JE Hirsch "An index to quantify an individual's scientific research output," http://xxx.arxiv.org/abs/physics/0508025