Abstract

Purpose – To identify the pros and the cons of Google Scholar.

Design/methodology/approach – Chronicles the recent history of the Google Scholar search engine from its inception in November 2004 and critiques it with regard to its merits and demerits.

Findings – Feels that there are massive content omissions presently but that, with future changes in its structure, Google Scholar will become an excellent free tool for scholarly information discovery and retrieval.

Originality/value – Presents a useful analysis for potential users of the Google Scholar site.

Keywords Reference services, Internet, Search engines

Paper type General review

The launch of Google Scholar – not surprisingly – drew much attention and praise, although not necessarily for the right reasons, both from the popular and the professional media. Google makes it very easy and free to find scholarly information about any topic – an important service for those who do not have access to the most appropriate fee-based indexing/abstracting databases which traditionally have helped in information discovery. Google Scholar goes beyond information discovery by leading qualifying users at subscribing libraries to the primary digital documents, and any users to the millions of open access (free) primary documents offered through mega-databases of preprint and reprint servers, as well as to the full text digital collections of several government agencies and research organizations. Google also deserves credit for introducing – albeit a bit belatedly – advanced options to refine the search process. On the negative side, the most important problem is that the crawlers of Google Scholar have not indexed millions of articles, even though they were let into the digital archives of most of the largest academic publishers and preprint servers and repositories. The stunning gaps give a false impression of the scholarly coverage of topics and lead to the omission of highly relevant articles by those who need more than just a few pertinent research documents. The rather enigmatic presentation of the results befuddles many users and the lack of any sort options frustrates the savvy searchers.

Pros

Google deserves credit for making the first step in providing a tool for discovering scholarly information, even though access may be limited to very short bibliographic citations of articles and conference papers in a sizable segment of Google Scholar. Still, many of the records have informative snippets from the full text for any users, and many also offer the abstracts of the articles for anyone. This free service alone is
roughly equivalent to what several of the traditional online indexing/abstracting databases have been providing for a hefty fee. The big plus is that Google searches the indexes created from the full text or part of the full text of the primary documents (even it is shows only a snippet of it), not merely the bibliographic records, abstracts and the subject terms (if assigned by the author or the publisher to the articles).

The crawlers of Google Scholar were let in the huge databases of the largest and most well-known scholarly publishers and university presses (such as IEEE, ACM, Macmillan, Wiley, University of Chicago); their digital hosts/facilitators (such as HighWire Press, MetaPress, Ingenta); societies and other scholarly organizations and government agencies (such as the American Physical Society, National Institute of Health, NOAA), and preprint/reprint servers (such as arXiv.org, Astrophysics Data System, RePEc, and CiteBase).

Patrons of libraries which have subscription to the digital archives of publishers are the greatest beneficiaries of Google Scholar, as with a single search they are lead to the digital full text versions of the articles and their supplements. This is particularly valuable for those libraries which have no federated search engines, and expect the patrons to repeat their searches by hopping from one publisher’s archive to the other, finding the query form and resubmitting the same query – which is not the prevailing attitude. In Google Scholar multiple database search is the default approach, unless the user specifies the publisher by using its URL in the site parameter, such as <tsunami site:ieee.org>. Once again, any user can see the bibliographic records, the abstract (if available with the paper), and/or the snippets of the context of the full text matching the query, and may order the document often at a much lower price than charged by some of the document delivery services.

The initial launch of Google Scholar in mid-November 2004 did not offer an advanced search template. In the review of the initial beta version (Jacsó, 2004), I voiced my disappointment that Google treats the highly structured records of scholarly articles the same way as the billions of unstructured web pages, even though the former have unambiguously and consistently tagged metadata identifying the title, the author, the journal name, the publication year and many other fields. Laudably, a month later the advanced template was introduced with good tips for providing additional search criteria to refine the searches (see Figure 1). Although some of them (such as the publication year) are not totally reliable, it is a good move by Google. So is the calculation and display of the “cited by” score, whose credence, however, should be established by disclosing the sources covered, and by vastly improving their currently unsystematic, unpredictable and disturbingly fragmentary coverage.

The coverage of Google is impressively broad and includes the most important scholarly publishers’ archives with the notable exception of Elsevier’s, the largest publisher. It is another question that the coverage of many archives is extremely shallow, which leads me to the cons.

Cons
The underlying problem with Google Scholar is that Google is as secretive about its coverage as the North Korean government about the famine in the country. There is no information about the publishers whose archive Google is allowed to search, let alone about the specific journals and the host sites covered by Google Scholar.
Another “feature” undisclosed by Google, but reported and well-illustrated among others by Gary Price specifically for Google Scholar (Price, 2004) is the fact that it limits the indexing of the collected files to the first 100-120K-bytes of the text (depending on the file type). The size of the majority of scholarly feature articles are close to or even exceed 1M-byte. If the search term occurs first beyond Google’s limit, the item would not be found. In fairness, AskJeeves has a similar limit, MSN is slightly more generous with a 150KB limit, and Yahoo! stops indexing at about 0.5MB (Sullivan, 2004).

My comparative search results in mid-November 2004 and on 1 January 2005 suggest that the content of Google Scholar has not been updated since its launch. This is not a major problem yet, but over time the staleness will become more prominent. Google has not disclosed yet how often it will be updated.

In this column I focus on the gaps still found through test searches in the first days of 2005 in the coverage of the most prominent archives covered by Google Scholar. The tests were done by limiting the subject search to specific site names which host the archives, such as <site:nature.com> for the Nature Publishing Group, <site:sciencemag.org> for Science magazine and <site:adsabs.harvard.edu> for the Astrophysics Data System, which is maintained by Harvard University among many other mirror sites. The site limiters are not always obvious, but from scanning search results the diligent user can figure them out. In some other cases a slightly different version of the domain name may yield somewhat different results, so one should proceed carefully (Jacsó, 2004). My special polysearch engine which I used for the test is available for anyone at www2.hawaii.edu/~jacso/scholarly/side-by-side2.htm

The test results were deeply disappointing in Google Scholar in light of what the native search engines of the sites retrieve for the same query. Through Google the search for “tsunami” in the title field limited to the site of the Nature Publishing Group
(NPG) yields a single record, while the native search engine finds eight items, all of them relevant.

Searching for the words “tsunami warning” in the full text using the two alternative tools shows an even bigger discrepancy; 17 items retrieved by the native search engine and only one by Google Scholar (see Figure 2).

For verification, the 16 additional hits were searched using the “allintitle” option in Google Scholar without any site limitations to see if the records may be available through other sources. One was found in ADS, two in the archive of the National Institute of Health (NIH). For one article there was one minimalist record found labeled as [CITATION]. This follow-up known-item search still resulted in an abysmal hit rate in Google Scholar (see Figure 3).
Figure 3. The ratio is far worse for full-text searching.
Apparently Google did not fully index the current eight-year collection, let alone the archive of *Nature* (which includes all the issues between 1987-1996) and the other 64 journals of NPG, all of which are hosted on the nature.com site. The native search engine at the NPG site finds nearly 87,000 records for items published in *Nature* alone between January 1987 and December 2004; Google Scholar finds only 13,700 records from the entire nature.com site. Using Google Scholar to search for the exact phrase “tsunami warning” in the *Nature* retrieves one hit for a 1993 article from the ADS database of Harvard and none from Nature’s archive (see Figure 4). The native search engine finds seven matching records.

The ADS record provides a link to the *Nature* archive, but the otherwise excellent ADS collection is not a substitute for *Nature*'s 18 years of digital coverage at its home site. In addition the test searches revealed that Google Scholar also has a puny coverage of ADS itself. The native search engine of the ADS database finds 32 records with the exact phrase “tsunami warning” in the abstract, while Google Scholar retrieves a mere nine records for the same query from the ADS database. It is quite telling about the shallowness of coverage that Google finds only 268,600 of the more than 4.1 million indexing/abstracting records in ADS.

A similar pattern is found when searching the ten-year archive of *Science* magazine (October 1995-December 2004) by the native search engine (nearly 40,000 records) versus Google Scholar with the site:scienccemag.org parameter (11,800 records). For the exact phrase “tsunami warning” Google Scholar retrieved a single record with this site limiter parameter, while the native search engine found three articles. Although Google Scholar does not always reproduce the same number of hits even when repeated within one hour intervals, these hit figures did show up consistently (see Figure 5). This was not the case when searching the site of the Proceedings of the National Academy of Sciences (PNAS), which retrieved 12,900 hits in late November, but 300 fewer records on 1 January 2005. I can only speculate that Google dynamically assigns the server to answer the query and the query-servers may not mirror their content exactly. The above three periodicals are among the most cited and most respected scholarly journals in their respective fields. If Google Scholar finds only 10-30 per cent of the records
which are available through using the sophisticated, still intuitive, native search engines, users would remain unaware of many potentially important articles.

These days, when scientists, administrators, politicians and financial experts need to find comprehensive and high quality scholarly information about the state of the art in tsunami warning systems to implement a feasible solution for the devastated Indian Ocean region, many will turn to Google Scholar to discover only a fragment of the scholarly literature. They also miss out on many scholarly papers which are open access (sometimes after an embargo of 3-12 months), as is the case with the poorly-covered papers published in the top-cited PNAS. Google has kept the beta label as a “shield” for some of its services for years The fact that Google Scholar is in beta version is not a good excuse for the massive content omissions. Google has used the beta “shield” for some of its services for over two years after their launch. Hopefully, Google Scholar will come out from its beta in a much shorter time, disclose the sources covered and fill the gaps to provide an excellent free tool for scholarly information discovery and retrieval.

References