NOTE

Equivalence of results from two citation analyses: Thomson ISI's Citation Index and Google's Scholar service

Daniel Pauly[1,*], Konstantinos I. Stergiou[2]

[1]Fisheries Center, University of British Columbia, 2202 Main Mall, Vancouver, British Columbia V6T 1Z4, Canada
[2]Department of Zoology, School of Biology, Aristotle University of Thessaloniki, PB. 134, 54124 Thessaloniki, Greece

ABSTRACT: Citation counts were performed across a wide range of disciplines using both the Thomson ISI files and Google Scholar, and shown to lead to essentially the same results, in spite of their different methods for identifying citing sources. This has strong implications for future citation analyses, and the many promotion, tenure and funding decisions based thereon, notably because ISI products are rather costly, while Google Scholar is free.

KEY WORDS: Citation analysis · Evaluations · Science policy

Since its introduction in the early 1960s, citation analysis has become a widespread evaluation tool (Lawrence 2003, King 2004). It was initially developed as a method for finding references other than by the then usual snowball method, by going backward through the references of citing papers. Its ability to move forward in time was soon used to identify highly referenced papers. This then allowed identification of highly cited scientists and research institutions, a transition accelerated by a series of contributions by Eugene Garfield and his associates at the Institute for Scientific Information (ISI). They demonstrated, based on ISI's unique database of painstakingly encoded references, that other indicators of scientific success (peer evaluation, membership in prestigious societies, prizes, etc.) strongly correlated with citation counts (Garfield 1977–1993). Since then, ISI-based evaluations have strongly affected academia (e.g. tenure decisions) and policy making (e.g. funding of scientists and universities), and led to international comparisons of scientific prowess (King 2004).

In November 2004, Google Inc. released the beta version of 'Google Scholar' (GS), which is based on software that identifies and gathers scientific papers from the web by identifying common formats of scientific papers and then extracts the title, authors, abstract, and references (Butler 2004). GS searches 'research publications such as journal articles, books, preprints and technical reports putting the most pertinent articles at the top of its searches' (Butler 2004). GS also 'searches abstracts from online archives such as PubMed and the NASA Astrophysics Data System and the complete text of physics preprints on the arXiv server' (Butler 2004). GS has agreements from 'almost all the 'major publishers'' to allow searches of the full text of their articles though GS declined to provide a list (Butler 2004). It is known that Elsevier, the largest scientific publisher, has refused to allow GS to search its texts. Nevertheless, GS, 'includes hits for more than a million Elsevier articles indexed as abstracts' (Butler 2004). Thus, GS is selective in what web based materials it searches.

We evaluated ISI and GS by comparing their citations of papers in mathematics, chemistry, physics, computing sciences, molecular biology, ecology, fisheries, oceanography, geosciences, economics, and psychology. Each discipline was represented by 3 authors, and each author was represented by 3 (i.e. high-,

medium-, and low-cited) articles (i.e. 99 articles). First, highly-cited authors we knew of from reading general literature were selected from both developed and developing countries. These were then complemented by randomly selected authors who referenced them. For both ISI and GS, citations to a given article were, in many cases, available in 'chunks', with the first chunk providing most of the citations, and the smaller chunks providing decreasingly smaller number of citations to what evidently was the same paper though its title or source may have exhibited differences in spelling or abbreviations used (see online supporting material: www.int-res.com/articles/suppl/E65_app.xls). Such cases were generally easy to spot and citations counts were summed. In addition, we included in our analysis 15 highly-cited articles (Garfield 1984). The 114 papers analyzed here were published from 1925 to 2004 in 75 journals, and were cited from 1 to over 100 000 times (the classic of Lowry et al. 1951). Belew (2005), in a similar analysis, uses 78 references for an unspecified number of disciplines/journals and a shorter time period (1977 to 2004).
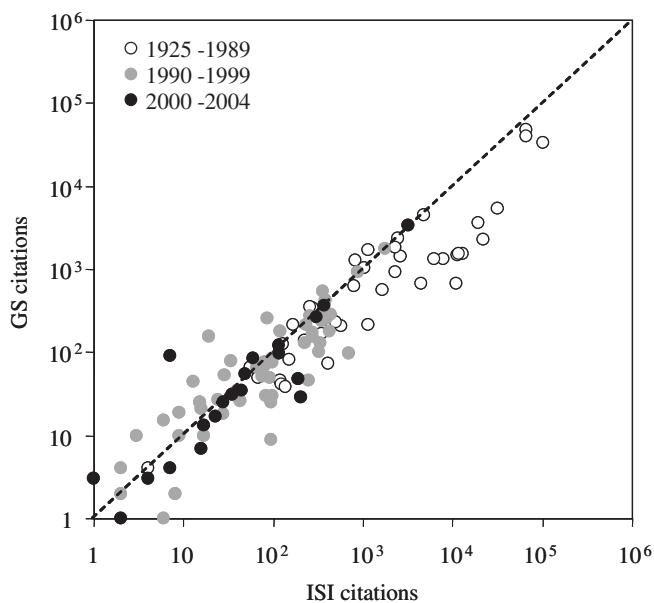

Fig. 1. Relationships between the citations in Thompson's ISI Citation Index (Web of Science, Full search, Cited ref search, 1970−today; access date: 5−25 September 2005) and Google Scholar (GS; same access dates) to 114 articles from 11 scientific disciplines, for 1925−2004 (see online supporting material). The data imply different regressions (lines not shown) for different time periods: (1) 1925−1989: (GS) = 0.454 (ISI) − 483.9 (r^2 = 0.83, p < 0.001, n = 42); (2) 1990−2004: (GS) = 0.991 (ISI) − 27.2 (r^2 = 0.95, p < 0.05, n = 72); (3) 2000−2004: (GS) = 1.026 (ISI) (r^2 = 0.994, p < 0.05, n = 20). For (2) and (3), slopes were not significantly different from unity (Student's t-test, t = −0.3511, p = 0.73 and t = 1.45, p = 0.16, respectively), indicating equivalence dotted line) between the 2 products

Fig. 1 presents our key results. For the period 1925 to 1989, the citations counts were proportional, but GS citations were less than half of ISI. This result was similar to Belew (2005), and was not unexpected. This is because most 'old' articles probably accumulated most of their citations relatively quickly and these citations were most probably from articles which, being 'old themselves', might not have yet been posted on the web. In contrast, for 1990 to 2004 and 2000 to 2004, not only were the citation counts proportional but the slopes were statistically indistinguishable from unity, suggesting that the citations in journals covered by ISI and picked up by GS were compensated by citations to other items on the web.

This is very surprising, given the character of the citing references in ISI and GS. ISI counts all the references of articles in several thousand pre-selected journals, while GS searches only scientific sources available on the web (Butler 2004). We expect GS's performance to improve for 'old' articles, as journals' back issues are posted on the web. Indeed, GS may gradually outperform ISI given its potentially broader base of citing articles.

Thus, GS can substitute for ISI, which so far has a monopoly (with the possible exception of Elsevier's very expensive search engine, Scopus). This has many implications relevant (as mentioned above), to science policy and to ethics, most emanating from the price differential between the costly ISI products and GS outputs, which presumably will continue to be free.

The price differential between ISI and GS might be particularly relevant for research and academic institutions in developing countries, and even modestly endowed institutions in developed countries (e.g. historically Black colleges and universities in the USA; Williams & Ashley 2004), which will be able to assess and document their scientific progress through GS at minimum cost. In addition, impact factors, or any other quantitative indicators, can in principle be computed using GS for any journal or other published item available online, not only for those listed in ISI. We hope that GS will make explicit routines available for such outputs.

We also think that free access to these data provided by GS offers an avenue for more transparency in tenure reviews, funding and other science policy issues, as it allows citation counts, and analyses based thereon, to be performed and duplicated by anyone. In this spirit, we also supply a spreadsheet as online supplementary material, which allows interested readers to check our data and inferences.

LITERATURE CITED

Belew RK (2005) Scientific impact quantity and quality: Analysis of two sources of bibliographic data. Available at: http://arxiv.org/abs/cs.IR/0504036 [accessed 19 Nov 2005]

Butler D (2004) Science searches shift up a gear as Google starts Scholar engine. Nature 432:423

Garfield E (1984) The 100 most-cited papers ever and how we select citation classics. Curr Cont 23:3–9

Garfield E (1977–1993) Essays of an information scientist, Vols 1–15. ISI Press, Philadelphia, PA

King DA (2004) The scientific impact of nations. Nature 430:311–316

Lawrence PA (2003) The politics of publication. Nature 422: 259–261

Lowry OH, Rosebrough NJ, Farr AI, Randall RJJ (1951) Protein measurement with the Folin phenol reagent. J Biol Chem 193:265–275

Williams J, Ashley D (2004) I'll find a way or make one: a tribute to historically Black colleges and universities. Amistad/HarperCollins, New York