Evidence Based Library and Information Practice

Evidence Summary

Use Google Scholar, Scopus and Web of Science for Comprehensive Citation Tracking

A review of:
Bakkalbasi, Nisa, Kathleen Bauer, Janis Glover and Lei Wang. "Three Options for Citation Tracking: Google Scholar, Scopus and Web of Science." Biomedical Digital Libraries 3.7 (2006).

Reviewed by:
Lorie A. Kloda
PhD student, School of Information Studies, McGill University
Montreal, Quebec, Canada
E-mail: lorie.kloda@mail.mcgill.ca

Received: 01 June 2007 **Accepted**: 12 July 2007

Abstract

Objective – To determine whether three competing citation tracking services result in differing citation counts for a known set of articles, and to assess the extent of any differences.

Design – Citation analysis, observational study.

Setting – Three citation tracking databases: Google Scholar, Scopus and Web of Science.

Subjects – Citations from eleven journals each from the disciplines of oncology and condensed matter physics for the years 1993 and 2003.

Methods – The researchers selected eleven journals each from the list of journals from *Journal Citation Reports* 2004 for the categories "Oncology" and "Condensed Matter Physics" using a systematic sampling technique to ensure journals with varying impact factors were included. All references from these 22 journals were retrieved for the years 1993 and 2003 by searching three databases: *Web of Science*, *INSPEC*, and *PubMed*. Only research articles were included for the purpose of the study. From these, a stratified random sample was created to proportionally represent the content of each journal (oncology 1993: 234 references, 2003: 259 references; condensed matter physics 1993: 358 references, 2003: 364 references). In November of 2005, citations counts were obtained for all articles from *Web of Science*, *Scopus* and *Google Scholar*. Due to the small sample size and skewed distribution of data, non-parametric

tests were conducted to determine whether significant differences existed between sets.

Main results – For 1993, mean citation counts were highest in *Web of Science* for both oncology (mean = 45.3, SD = 77.4) and condensed matter physics (mean = 22.5, SD = 32.5). For 2003, mean citation counts were higher in *Scopus* for oncology (mean = 8.9, SD = 12.0), and in *Web of Science* for condensed matter physics (mean = 3.0, SD = 4.0). There was not enough data for the set of citations from *Scopus* for condensed matter physics for 1993 and it was therefore excluded from analysis. A Friedman test to measure for differences between all remaining groups suggested a significant difference existed, and so pairwise post-hoc comparisons were performed. The Wilcoxon Signed Ranked tests demonstrated significant differences "in citation counts between all pairs ($p < 0.001$) except between *Google Scholar* and *Scopus* for CM physics 2003 ($p = 0.119$)."

The study also looked at the number of unique references from each database, as well as the proportion of overlap for the 2003 citations. In the area of oncology, there was found to be 31% overlap between databases, with *Google Scholar* including the most unique references (13%), followed by *Scopus* (12%) and *Web of Science* (7%). For condensed matter physics, the overlap was lower at 21% and the largest number of unique references was found in *Web of Science* (21%), with *Google Scholar* next largest (17%) and Scopus the least (9%). Citing references from *Google Scholar* were found to originate from not only journals, but online archives, academic repositories, government and non-government white papers and reports, commercial organizations, as well as other sources.

Conclusion – The study does not confirm the authors' hypothesis that differing scholarly coverage would result in different citation counts from the three databases. While there were significant differences in mean citation rates between all pairs of databases except for *Google Scholar* and *Scopus* in condensed matter physics for 2003, no one database performed better overall. Different databases performed better for different subjects, as well as for different years, especially *Scopus*, which only includes references starting in 1996. The results of this study suggest that the best citation database will depend on the years being searched as well as the subject area. For a complete picture of citation behaviour, the authors suggest all three be used.

Commentary

This study makes a contribution to current research on citation databases similar to Jacso, Yang and Meho. (For a more complete list of recent research on the subject, see Schroeder.) It adds a unique aspect by investigating the citation counts for two very specific subject areas – oncology and condensed matter physics – that present different publishing patterns. The citation analysis method used is appropriate for verifying the study's hypothesis, however there are some issues concerning the choice of years and databases. *Web of Science* is often considered the premier database for citation searching; *Scopus* and *Google Scholar* are still in development. Certainly, the citation sets for 1993 reflect this, since *Scopus* has yet to add backfiles prior to 1996. Since this study was published, *Scopus* has added over 800 new journal titles to its database, a number which would certainly impact the results if this study were repeated. *Google Scholar* remains the newest, and least transparent of the databases, providing, for example, no information as to which publication dates are covered.

Other methodological issues include the small size of the sample, and the limitation of the subject areas, making it difficult to statistically generalise the findings. The authors do not take into account publications other than research articles. The sample is drawn from Thomson's *Journal Citation Reports*, giving *Web of Science* the advantage, as all these journals are indexed within it. Journal self-citations can sometimes be quite high, and a paper published in a *Web of Science*-indexed journal, statistically speaking, will have a greater chance of being cited than a paper that is published in a journal which is not indexed in the database. For the citation search itself, the authors do not explain how *Google Scholar* was searched, as this database is not searchable in the same way as the other, more traditional citation tracking tools. The value of this research could have been enhanced by including the researchers' search strategy either in the text, or in an appendix.

Although more research needs to be done, it is clear from the results of this study that no one database will suffice for calculating an article's citation counts. Librarians would be well advised to use all three when possible and to educate users to do the same. In addition, it is recommended that all three be used for locating research when an alternative to traditional subject searching is desired. Research looking at third-party software to combine data from these databases, or others that include citation information, is needed (see Yang and Meho for an example).

Citation tracking is more popular than ever. As more providers offer tools for performing citation searches, librarians will need to be familiar not only with the best tools for the job, but with the surrounding issues as well. The proliferation of these databases indicates an increased importance in citation tracking, not just for locating related research, but for measuring academic output and performance. As information professionals, we need to educate researchers about the pitfalls of using the impact factor and citation counts in decision making for hiring, promotion, tenure and funding. Questions remain about the validity of using these measures as indicators of quality of research output, and alternatives such as the Eigenfactor (Bergstrom) and the h-index (Hirsch) should be investigated.

Works Cited

Bergstrom, C. "Eigenfactor: Measuring the Value and Prestige of Scholarly Journals." College & Research Libraries News 68.5 (2007): 314-316.

Hirsch, J.E. "An Index to Quantify an Individual's Scientific Research Output." Proceedings of the National Academy of Sciences 102.46 (2005):16569-16572.

Jacso, Peter. "As We May Search – Comparison of Major Features of the *Web of Science*, *Scopus*, and *Google Scholar* Citation-based and Citation-enhanced Databases." Current Science 89.9 (2005): 1537-1547.

Schroeder, Robert. "Pointing Users Toward Citation Searching: Using Google Scholar and Web of Science." portal: Libraries and the Academy 7.2 (2007): 243-248.

"Scopus Qualifies More than 800 New Titles to Be Added to Its Database." Press release. Amsterdam. 30 Apr. 2007. 13 July 2007 <http://www.info.scopus.com/news/press/pr_300407.asp>.

Yang, Kiduk and Lokman I. Meho. "Citation Analysis: A Comparison of Google Scholar, Scopus, and Web of Science." In Grove, Andrew, ed. Proc. of the 69th Annual Meeting of the American Society for Information Science and Technology (ASIST) 43, Austin, 2006.