CiteSearch: Next-generation Citation Analysis

Kiduk Yang
Indiana University
1320 E. 10th Street, LI013
Bloomington, IN 47405
1-812-855-2793
kiyang@indiana.edu

Lokman Meho
Indiana University
1320 E. 10th Street, LI005C
Bloomington, IN 47405
1-812-855-2323
meho@indiana.edu

ABSTRACT

The coverage of citations in citation databases of today is disjoint and incomplete, which can result in conflicting quality assessment outcomes across different data sources. Fusion approach to quality assessment that employs a range of citation-based methods to analyze data from multiple sources is one way to address this limitation. The paper discusses a citation analysis pilot study that measured the impact of scholarly publications based on the data mined from Web of Science, Scopus, and Google Scholar.

Categories and Subject Descriptors

H.3.7 [**Information Storage and Retrieval**]: Online Information Services – *Web-based services*

General Terms

Measurement, Performance, Experimentation

Keywords

Citation Analysis, Bibliometrics, Quality Assessment

1. INTRODUCTION

The basic assumption underlying citation analysis is that citations are a way of giving credit to and recognizing the value, quality, or significance of an author's work [2, 12]. While the proponents have reported the validity of using citation counts for research assessments [1, 6], critics claim that citation analysis has serious limitations in both data and methodology [7, 10]. The problems reported in literature point to two fundamental shortcomings with the typical citation analysis approach. First, conventional citation analysis methods yield one-dimensional and sometimes misleading evaluation as a result of not taking into account differences in citation quality, not filtering out citation noise such as self-citations, and not considering non-numeric aspects of citations such as language, culture, and time. Second, the coverage of citations in citation databases of today is disjoint and incomplete, which can result in conflicting quality assessment outcomes across different data sources.

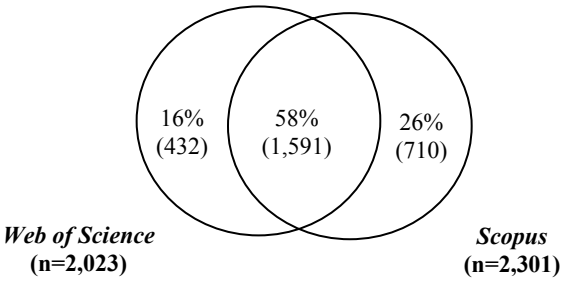
JCDL'07, June 18–23, 2007, Vancouver, British Columbia, Canada.

To address these limitations and produce a reliable and efficient indicator for assessing the relative impact and quality of scholarly publications, the *CiteSearch* project at Indiana University is developing a multi-faceted fusion approach to information quality assessment that employs a range of citation-based methods to analyze data from multiple sources. The paper discusses the *CiteSearch* pilot study that measured the impact of scholarly publications based on the data mined from Web of Science, Scopus, and Google Scholar.

2. *CITESEARCH* STUDY

In the pilot study, we identified citations to each of the 1,093 items published by the 15 faculty members of Indiana University School of Library and Information Science (SLIS). All *Scopus* and *Web of Science* data were manually collected and processed twice in October 2005 and again (for accuracy and updating purposes) in March 2006. *Google Scholar* data were harvested by *CiteSearch* system in March 2006.

Figure 1. Distribution of unique and overlapping citations in *Web of Science* and *Scopus* (N=2,733)

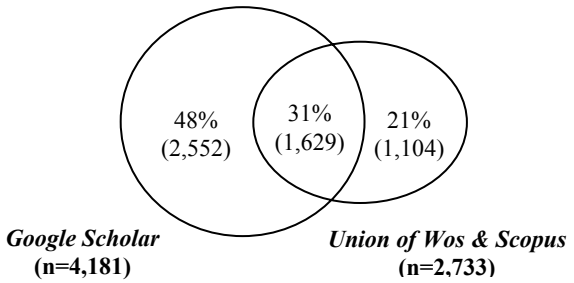


The result of our study shows Scopus to have 278 (14%) more citations than Web of Science, suggesting that Scopus provides more comprehensive coverage of the LIS literature than Web of Science. Further analysis of the data shows that combining citations from Scopus and Web of Science increases the number of citations of SLIS as a whole by 35%. Data also show that the percentage of increase in citation counts for individual faculty members varies considerably depending on their research areas, ranging from 5% to 99%. These findings not only imply that certain subject areas will benefit more than others from using both Scopus and Web of Science to identify relevant citations, they also suggest that to generate accurate citation counts for faculty members, and by extension schools, and to accurately compare them to one another, a researcher must use both databases. The importance of using Scopus in addition to Web of Science is further evidenced by the fact that the relative ranking of faculty

members changes in eight out of 15 cases. In addition, Scopus retrieves considerably more citations from refereed conference papers than Web of Science (359 vs. 229). 54% of citations from conference papers are uniquely found in Scopus in comparison to only 28% in Web of Science (20% of citations from conference papers are found in both databases). This could have significant implications for citation analyses and the evaluation of individual scholars, especially when those evaluated include authors who use conferences as a main channel of scholarly communication.

The findings suggest that many of the previous studies that used Web of Science exclusively to generate citation data to evaluate and/or rank scholars, journals, programs, and so on have been based on skewed and incomplete data and may have resulted in inaccurate assessments and imprecise rankings. Given the low overlap in citations between the two databases, the findings further suggest that the use of Scopus in addition to Web of Science may have significant implications on the h-index scores of authors and journals [3, 5], and journal impact factors [4, 9].

Figure 6. Distribution of unique and overlapping citations in Google Scholar and *WoS_Scopus* (N=5,285)



In contrast to *Web of Science* and *Scopus*, which index citations mainly from journal articles and conference papers, citations found through *Google Scholar* come from many different types of documents. Results show that *Google Scholar* identifies 1,448 (53%) more citations than *Web of Science* and *Scopus* combined (4,181 vs. 2,733) and combining citations from all three sources increases the number of citations by 93% (from 2,733 to 5,285 citations). In other words, one would miss over 93% of relevant citations if searching were limited to *Web of Science* and *Scopus*.

Despite the large increase in citation counts, adding *Google Scholar*'s unique citations data to those of *Web of Science* and *Scopus* does not significantly alter the relative ranking of faculty members—Spearman Rank Order correlation coefficient = 0.976 at 0.001 level. *Google Scholar* also misses 1,104 (40%) of the 2,733 citations found by *Web of Science* and *Scopus* (Figure 2). This is strikingly high, especially given the fact that virtually all citations from *Web of Science* and *Scopus* come from refereed and/or reputable sources.

Given the fact that *Google Scholar* is so cumbersome to use, misses a significant number of citations from refereed sources, and has little or no influence on the relative rankings of scholars, one could conclude that, as far as LIS is concerned, *Google Scholar* is superfluous, especially when the focus of a study is on citations in refereed journals and conference proceedings and when both *Web of Science* and *Scopus* are used to generate citation counts for assessing and comparing scholars, journals, and academic departments to one another. *Google Scholar*, however, could be very useful for individual scholars preparing for tenure and promotion and/or for comparative citation analysis studies that attempt to map or visualize scholarly networks [11, 13] and/or those studies that try to show one's international impact [8]. It should be emphasized here that *Google Scholar* uncovers many unique citations (2,552 or 48% of the 5,285 citations found in all three sources) that could be very useful in such studies or for such purposes (Figure 2).

REFERENCES

[1] Adkins, D., & Budd, J. (2006). Scholarly productivity of U.S. LIS faculty. *Library & Information Science Research*, 28(3), 374-389.

[2] Cronin, B. (1984). *The citation process: The role and significance of citations in scientific communication*. London: Taylor Graham.

[3] Cronin, B., & Meho, L. (2006). Using the *h*-index to rank influential information scientists. *Journal of the American Society for Information Science and Technology*, 57(9), 1275-1278.

[4] Garfield, E. (1996). How can impact factors be improved? *British Medical Journal*, 313 (7054), 411-413.

[5] Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, 102(46), 16569-16572. Retrieved February 15, 2006, from http://www.pnas.org/cgi/reprint/102/46/16569.

[6] Holmes, A., & Oppenheim, C. (2001). Use of citation analysis to predict the outcome of the 2001 Research Assessment Exercise for Unit of Assessment (UoA) 61: Library and Information Management. *Information Research*, 6(2). Retrieved June 15, 2005, from http://informationr.net/ir/6-2/paper103.html.

[7] MacRoberts, M.H., & MacRoberts, B.R. (1996). Problems of citation analysis. *Scientometrics*, 36(3), 435-444.

[8] Nisonger, T.E. (2004). Citation autobiography: An investigation of ISI database coverage in determining author citedness. *College & Research Libraries*, 65(2), 152-163.

[9] Nisonger, T. E. (2004). The benefits and drawbacks of impact factor for journal collection management in libraries. *The Serials Librarian*, 47(1-2), 57-75.

[10] Seglen, P.O. (1998). Citation rates and journal impact factors are not suitable for evaluation of research. *Acta Orthopaedica Scandinavica*, 69(3), 224-229.

[11] Small, H. (1999). Visualizing science by citation mapping. *Journal of the American Society for Information Science*, 50(9), 799-813.

[12] van Raan, A.F.J. (1996). Advanced bibliometric methods as quantitative core of peer-review based evaluation and foresight exercises. *Scientometrics*, 36(3), 397-420.

[13] White, H. D., & McCain, K. W. (1997). Visualization of Literatures. *Annual Review of Information Science and Technology*, 32, 99-168.