Document type assignment accuracy in citation index data sources

Paul Donner

donner@forschungsinfo.de
Institute for Research Information and Quality Assurance (iFQ), Schützenstr. 6a, 10117 Berlin (Germany)

Introduction

The observed citation counts of publications can be divided by the average of a reference set of similar publications in order to get a relative impact measure. It is customary to define the reference set by publication date, scientific discipline and document type. Different document types (DT) have very different citation distributions, leading to very different results in calculations of indicators when separating reference sets by DT and disregarding this kind of normalization (Sirtes, 2012). Thus, when computing relative impact, the correctness of the assignment of document types to publications is crucial. The correctness of DT assignment in citation indexes has been called into question by studies of van Leeuwen et al. (2007), drawing attention to the treatment of letters and 'research letters' from medical journals as the same type in Web of Science and by Harzing (2003), illustrating how WoS is using some highly questionable assignment criteria. In this contribution DT assignments in WoS (Thomson Reuters, 2013) and Scopus (Elsevier, 2014) by their respective staff are compared to those of the publishers.

Methods and data

For this study data licenced from Thomson Reuters Web of Science and Elsevier Scopus and loaded into SQL databases was used. The databases are part of the infrastructure of the German Competence Centre for Bibliometrics project. Random samples of document identifiers were drawn from the WoS records, stratified by DT as assigned in WoS, restricted to items published in journals. Subsamples of the document types 'article', 'review' and 'letter', as well as of records not assigned to any of those three types (here called 'other') were taken. This follows the convention of distinguishing between 'citable items' and others. They were linked to the Scopus records detailing the same documents using DOIs. It follows that only documents with a DOI are used. In the resulting sample table, only the WoS and Scopus document identifiers and the DOI are saved in a row. The rows were randomized.

To each sample record, bibliographic description data comprised of article title, first author family name and initials, publication year, journal name, volume and issue were queried from the WoS data and saved along with record IDs into a separate table. Student assistants were tasked to search for the article abstract web pages online using the bibliographic information to query Google Scholar and web search. On the individual article web page of the journal, they were instructed to find the officially assigned document type, if specified, and code it as article, letter, review, other or not found. If no type was stated but it was clearly deducible from the abstract or title, this was also accepted.

A sample of 528 publications was analyzed so far, on which the following provisional results are based. For a further 90 publications, no certain DT assignment was possible. Found (true) DT and Scopus/WoS DT were tabulated and classified as true/false positive/negative. From those counts precision and recall were computed for each DT and combined precision and recall as weighted by DT occurrence frequency in the databases. The effect of false DT assignment on publication normalized citation score is measured in percent deviation.

Results

The results depicted in Fig. 1 show that in both citation indexes the accuracy of correct DT assignment is quite poor. WoS gives the correct DT in about 72%, Scopus in about 80% of cases (as weighted by shares of DT in the databases). On average WoS finds about 81% of publications of a given DT while Scopus will return about 73%. Error bars for the DT specific results are 95% posterior probability Bayesian credible intervals for the binomial proportion, using a flat beta prior with both shape parameters set to 1.

These findings necessarily have an adverse effect on the mean field/DT/year specific expected citation rates used as reference standards in obtaining normalized publication level citation scores. To give an idea of the magnitude of this effect, the normalized article citation score (3-year citation window) for publications that were assigned an incorrect DT in WoS was calculated following Waltman et al. (2011).

The differences between incorrect and correct score in percent of the correct score are plotted as a histogram in Fig. 2. Publications with zero citations are not used (N_0=34), since no difference could manifest.

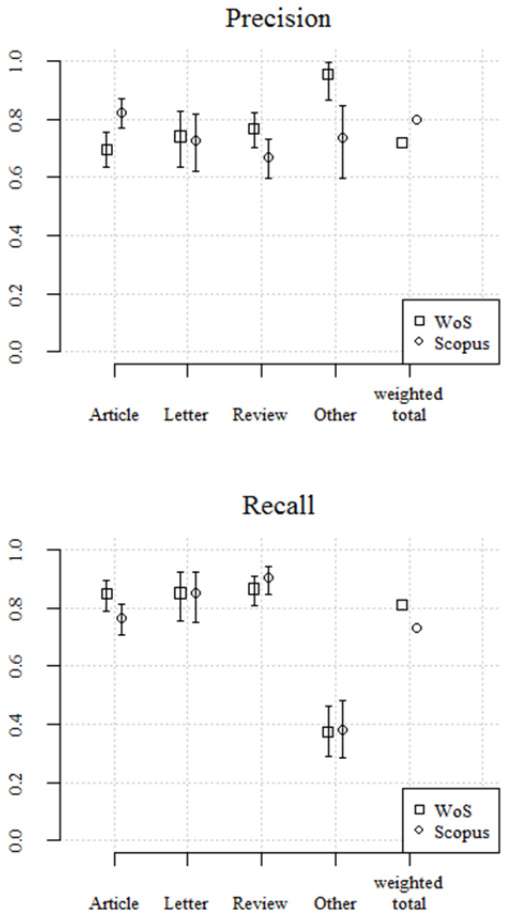

Figure 1. Precision and recall per document type in WoS and Scopus (N=528).

Conclusion

Document type assignment is unreliable in both Web of Science and Scopus and will cause large errors in publications' normalized citation scores and consequently derived indicators such as field-normalized mean citation rate.

References

Elsevier B.V. (2014). Scopus Content Coverage Guide 07.14. [accessed 2015/02/06] http://www.elsevier.com/__data/assets/pdf_file/0011/242489/Content-Coverage-Guide.pdf

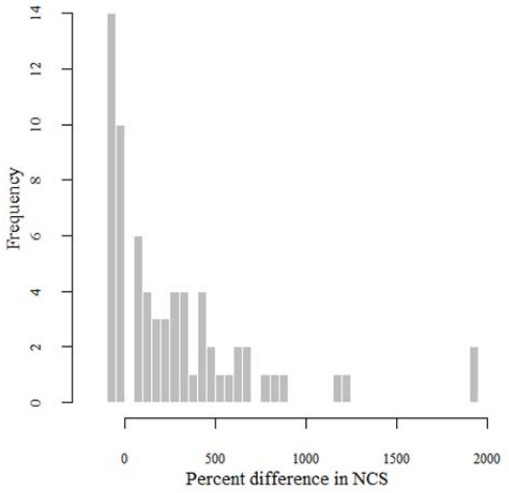

Figure 2. Percent difference in normalized citation score per document for those with wrong DT assignment in WoS (N=68).

Harzing, A.W. (2013). Document categories in the ISI Web of Knowledge: Misunderstanding the Social Sciences? *Scientometrics*, 93(1), 23-34.

Sirtes, D. (2012) How (dis-) similar are Different Citation Normalizations and the Fractional Citation Indicator? (And How it can be Improved). É. Archambault, Y. Gingras & V. Larivière (Eds.), *Proceedings of 17th International Conference on Science and Technology Indicators (STI)*, Montréal: Science-Metrix and OST, 894-896.

Thomson Reuters (2013). Web of Science® Help. Searching the Document Type Field. [accessed 2015/02/06], http://images.webofknowledge.com/WOKRS59B4/help/WOS/hs_document_type.html

Van Leeuwen, T. N., Van Der Wurff, L. J., & De Craen, A. J. M. (2007). Classification of "research letters" in general medical journals and its consequences in bibliometric research evaluation processes. *Research Evaluation*, 16(1), 59-63.

Waltman, L., van Eck, N. J., van Leeuwen, T. N., Visser, M. S., & van Raan, A. F. (2011). Towards a new crown indicator: Some theoretical considerations. *Journal of Informetrics*, 5(1), 37-47.