Essays

The Subject Sameness Index: do author-provided keywords extracted from bibliographic databases provide comparable data?

Frank-Thorsten Krell
Denver Museum of Nature & Science, 2001 Colorado Boulevard, Denver, Colorado 80205-5798, USA; frank.krell@dmns.org

The newly proposed Subject Sameness Index[1] (SSI) is an interesting suggestion which is worth exploring. It is a novel indicator that can be used to describe the performance of researchers, indicating the broadness of one's research, or 'field mobility' as it is called in the scientometric literature.[2] Field mobility, or extent of subject sameness of a scientist's publications, is certainly an attribute of major interest with search or promotion panels, but should not be misinterpreted. Just as the frequently used citation metrics do not indicate the quality of research output,[3] keywords likewise fail in this respect. Broader focus does seem to be positively correlated with higher productivity,[4] but does not necessarily go along with higher quality of research. It is a possible consequence of increasing experience and maturity of a researcher. A simple, easy to calculate index determining field mobility such as the proposed SSI, would be a welcome tool if widely applicable and reliably fulfilling its function. Let's explore whether Tirgar *et al*'s method[1] to quantify field mobility or subject sameness with a simple index can lead to robust results.

The poverty of citation databases

All available commercial bibliographical databases are incomplete.[5] Not the productivity of the researchers determined the data used by Tirgar *et al*,[1] but the selective inclusion of their papers in the Web of Science® (WoS) database. Commercial bibliographical databases, particularly WoS, are biased in favour of English language journals,[6] preferentially from the Western world. This probably put the Iranian researchers, used by Tirgar *et al*[1] as a sample group, at a disadvantage. Indeed, a decade ago Moin *et al*[7] mentioned that only three of Iran's 247 scientific journals were registered with ISI® (WoS). It is unlikely that this situation has changed dramatically. It is easily possible that WoS contains less than half of the overall publication output of the Iranian scholars studied. Particularly with authors from non-Anglophone, non-Western countries, the SSI would be more reliably determined by using the complete set of papers of those authors.

Are author-provided keywords a reliable source for determining sameness?

The consideration of keywords for determining subject sameness is problematic if this data source is uncritically utilised. Problems arise from differing keyword policies of scientific journals. Author-provided keywords are common in the biomedical literature, but not ubiquitous as they are lacking, for example, from articles published in *Nature*, *Science*, or *Cell*. Journals that require keywords generally restrict the number of keywords. Although the majority of journals, particularly in biomedical disciplines, request a maximum of four to six keywords, others, such as *BMC Genomics* or *European Journal of Cancer* allow up to ten. This discrepancy is likely to confound any comparative metrics based on keywords. A larger number of keywords allows for additional fringe terms that might not necessarily represent the core topic of the article, but covers its contents more completely. Such more complete coverage is likely to permit overlap with keywords of articles of a similar topic but different focus. Articles in journals requiring a low number of keywords do overlap only with very similar papers.

In some disciplines, such as ecology and natural history, a considerable number of journals require keywords to be different from words used in article titles, eg *Basic and Applied Ecology*, *Caribbean Journal of Science*, *Evolution*, *Journal of Animal Ecology*, or *Landscape Ecology*. Without extracting keywords from article titles and adding them to the author-provided keywords, subject sameness cannot be determined in those fields. In co-word analysis, another keyword based analysis, additional indexing is generally done, be it of title words or the full text, to achieve a useful data basis.[8]

Conclusion

The proposed SSI, being easy to calculate, is an attractive scientometric tool with various applications. Before we can use it as a trusted metric, it needs to be explored whether and how incongruent keyword policies of journals pose a problem for its comparability. For small data sets, only papers of journals with the same keyword policies and numbers are comparable. Large datasets might 'equal out' slight differences in keyword numbers and policies, but analysing single authors might not allow for large enough datasets. As with co-word analysis,[8] we might end up employing additional indexing in order to create comparable data for more robust results.

References

1 Tirgar A, Yaminfirooz, M, Ahangar HG. Subject Sameness Index: a new scientometric indicator. *European Science Editing* 2013;39(1):3-4.

2 Hellsten I, Lambiotte R, Scharnhorst A, Ausloos M. Self-citation, co-authorships and keywords: A new approach to scientists' field mobility? *Scientometrics* 2007;72(3):469–486.

3 Krell F-T. The Journal Impact Factor as a performance indicator. *European Science Editing* 2012;38(1):3–6.

4 Heeringen A van, Dijkwel PA. The relationship between age, mobility and scientific productivity. Part I. *Scientometrics* 1987;11(5–6):267–280.

5 Krell F-T. The poverty of citation databases: (...). *BioScience* 2009;59(1):6–7.

6 Raan AFJ van, Leeuwen TN van, Visser MS. Severe language effects in university rankings: (...) *Scientometrics* 2011;88:495–498.

7 Moin M, Mahmoudi M, Rezaei N. Scientific output of Iran at the threshold of the 21st century. *Scientometrics* 2005;62(2):239–248.

8 Ding Y, Chowdhury GG, Foo S. Bibliometric cartography of information retrieval research by using co-word analysis. *Information Processing and Management* 2001;37:817–842.