Research Trends

Volume 1 | Issue 30 Article 8

9-1-2012

The use of big datasets in bibliometric research

Henk F. Moed Dr

Section 7:
The use of Big Datasets in bibliometric research

Henk F. Moed
Senior Scientific Advisor, Elsevier,
Amsterdam, The Netherlands

Introduction

Due to the increasing importance of scientific research for economic progress and competitiveness, and to new developments in information and communication technologies (ICT), the fields of bibliometrics and research assessment are rapidly developing. A few major trends can be identified:

- An increase in actual use of bibliometric data and indicators in research assessment;
- A strong proliferation of bibliometric databases and data-analytical tools; for instance, in the emergence of a range of journal subject classification systems and key words mapping tools;
- Indicators are becoming more and more sophisticated and fit-to-purpose; new approaches reveal that bibliometrics concerns much more than assessing individuals on the basis of journal impact factors;
- There is an increasing interest in measuring the effects of the use of bibliometric indicators upon the behavior of researchers, journal editors and publishers;
- Researchers, research evaluators and policy officials place an emphasis on the societal impact of research, such as its technological value or its contribution to the enlightenment of the general public;
- Last but not least, more and more projects aim to create and analyze large datasets by combining multiple datasets.

This article deals with the last trend mentioned and focuses on demonstrating which datasets are currently being combined by research groups in the field. It also discusses the aspects and research questions that could be answered using these large datasets. An overview is given in Table 1.

Examples

Downloads versus citations

For a definition of "usage" or "downloads" analysis and its context the reader is referred to a previous RT article on this topic (1). Figure 1 relates to journals included in ScienceDirect, Elsevier's full text article database. For each journal the average citation impact per article was calculated (generated in the third year after publication date), as well as the average number of downloads in full text format per article (carried out in the year of publication of the articles). Journals were grouped into disciplines; the horizontal axis indicates the number of journals in a discipline. In each discipline the Pearson correlation coefficient between a journal's downloads and its citations was calculated, and plotted on the vertical axis.

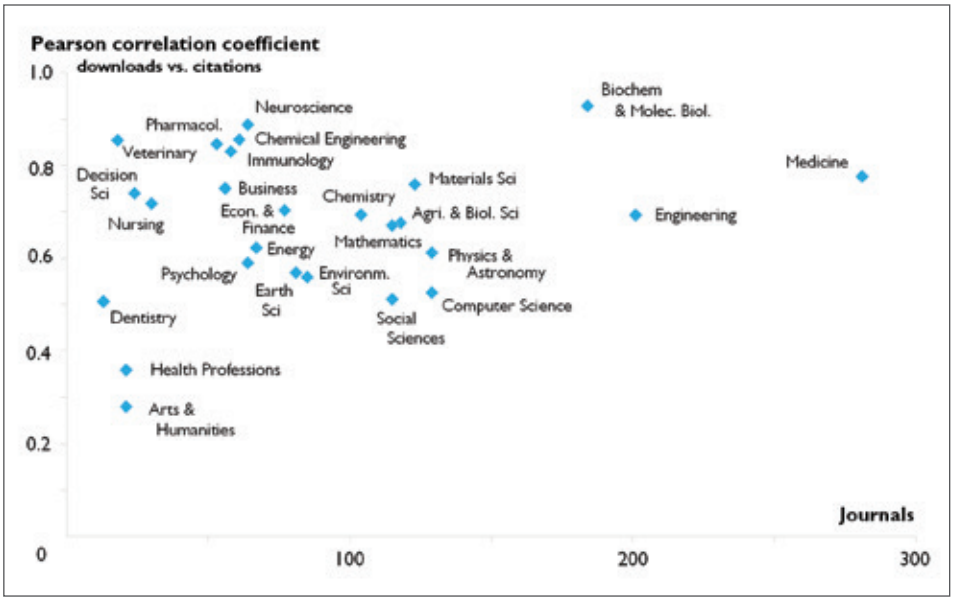

Figure 1: Downloads versus citations for journals in ScienceDirect.

Combined datasets	Studied phenomena	Typical research questions
Citation indexes and usage log files of full text publication archives	Downloads versus citations; distinct phases in the process of processing scientific information	What do downloads of full text articles measure? To what extent do downloads and citations correlate?
Citation indexes and patent databases	Linkages between science and technology (the science–technology interface)	What is the technological impact of a scientific research finding or field?
Citation indexes and scholarly book indexes	The role of books in scholarly communication; research productivity taking scholarly book output into account	How important are books in the various scientific disciplines, how do journals and books interrelate, and what are the most important books publishers?
Citation indexes (or publication databases) and OECD national statistics	Research input or capacity; evolution of the number of active researchers in a country and the phase of their career	How many researchers enter and/or move out of a national research system in a particular year?
Citation indexes and full text article databases	The context of citations; sentiment analysis of the scientific-scholarly literature	In what ways can one objectively characterize citation contexts? And identify implicit citations to documents or concepts?

Table 1: Compound Big Datasets and their objects of study.

Figure 1 reveals large differences in the degree of correlation between downloads and citations between disciplines. For instance, in Biochemistry and Molecular Biology the correlation is above 0.9, whereas in Dentistry, Social sciences, Health Professions, Arts and Humanities it is equal to or less than 0.5.

The interpretation of these findings is somewhat unclear. One hypothesis is based on the distinction between authors and readers. In highly specialized subject fields these populations largely overlap, whereas in fields with a more direct societal impact, the readers' population may consist mainly of professionals or even the general public who do not regularly publish articles. The hypothesis proposes that in the latter type of fields the correlation between downloads and citations is lower than in the first. Additional research, also conducted at the level of individual articles, is needed to further examine this hypothesis.

Patents and scientific articles

Earlier this year, Research Trends also published an article analyzing patent citations to journal articles, in order to measure the technological impact of research (2). The analysis focused on a subject field in the social sciences. It examined the characteristics of research articles published in Library Science journals and the manner by which they are cited in patents. Library science articles were found to be well cited in patents. The articles cited feature information retrieval and indexing, and information and documents management systems which pertain to electronic and digital libraries development. The citing patents focus on electronic information administration, navigation, and products and services management in commercial systems. Interestingly, the time span between the scientific invention and its use in technology may be up to 10 years. This finding illustrates the time delays one has to take into account when trying to measure technological or societal impact of scientific research. For an overview of this way of using patent citations, see (3).

Scopus author data versus OECD "input" statistics

Scopus, Elsevier's scientific literature database, containing meta-data of scientific publications published by more than 5,000 publishers in 18,000 titles, has implemented unique features that enable one to obtain an estimate of the number of active – i.e., publishing – authors in a particular year, country, and/or research domain, and also to track the "institutional" career of a researcher, providing information on the institutions in which a researcher has worked during his or her career. Research Trends issues 26 and 27 contained two articles by Andrew Plume presenting a first analysis of migration or brain circulation patterns in the database (4) (5).

Data accuracy and validation is also a relevant issue in this case. One way to validate author data is by comparing outcomes per country with statistics on the number of full time equivalents spent on research in the various institutional sectors, obtained from questionnaires and published by the OECD.

Country	Germany	UK	Italy	The Netherlands
OECD number of FTE Research 2007 (all sectors)	290,800	254,600	93,000	49,700
OECD number of FTE Research 2007 (Higher Education & Government sector)	116,600	159,100	56,200	23,800
Number of Publishing authors in Scopus	150,400	154,600	113,100	46,300
Ratio number of authors / Number of FTE Research (all Sectors)	0.52	0.61	1.22	0.93
Ratio number of authors / Number of FTE Research (Higher Education & Government sector)	1.29	0.97	2.01	1.95

Table 2: OECD and Scopus based "input" statistics for 4 European countries.

Table 2 presents statistics for 4 countries. Rather than comparing absolute numbers, it is interesting to calculate the ratios in the last two rows of the table. It is striking that these ratios differ substantially between countries. They are much higher for the Netherlands and Italy than they are for Germany and UK. This outcome points first of all towards the need to further validate Scopus-based numbers of active researchers. On the other hand, it also raises the question whether the various countries have applied the same definition of FTE research time in their surveys.

Books and journals

Scientific-scholarly books are generally considered as important written communication media, especially in social sciences and humanities. There is an increasing interest in studies of the function and quality of books and book publishers in the various domains of science and human scholarship. Thomson Reuters has launched its Book Citation Index. The Google Books project aims to digitalize millions of books, including many scientific-scholarly ones. Expanding a primarily journal-based citation index with scholarly book sources has two advantages. Not only is the set of source publications expanded with relevant sources, but the enormous reservoir of cited references given in journal articles to book items is used more efficiently.

Citations and full texts

The availability of full text research articles in electronic format gives us the opportunity to conduct textual analyses of all of an article's content – not just the meta-data extracted by indexing databases. The citation contexts can be analyzed linguistically, and sentiment analyses can be conducted to reveal how the citing author appreciates a cited work. Henry Small and Richard Klavans used citation context analysis as an additional tool for the identification of scientific breakthroughs (6). In one of its next issues Research Trends will publish an article on a detailed citation context analysis in one particular journal focusing on cross-disciplinary citations.

Concluding remarks

The overview above is not complete, and many important contributions to the analysis of big, compound bibliometric datasets were not mentioned in this paper. But the examples presented above illustrate the theoretical and practical relevance of combining bibliometric, or, more generally, statistical datasets, show how this can be done, and indicate which issues a big, compound, bibliometric dataset enables us to address.

References:

1. Lendi, S. & Huggett, S. (2012) "Usage: an alternative way to evaluate research", Research Trends, No. 28 http://www.researchtrends.com/issue28-may-2012/usage-an-alternative-way-to-evaluate-research/

2. Halevi, G. & Moed, H.F. (2012) "Patenting Library Science Research Assets", Research Trends, No. 27 http://www.researchtrends.com/issue-27-march-2012/patenting-library-science-research-assets/

3. Breschi, S. & Lissoni, F. (2004) "Knowledge Networks from Patent Data. In: Moed, H.F., Glänzel, W., and Schmoch, U. (eds.). Handbook of quantitative science and technology research. The use of publication and patent statistics in studies of S&T systems. Dordrecht (the Netherlands): Kluwer Academic Publishers, 613-644.

4. Plume, A. (2012) "The evolution of brain drain and its measurement: Part I", Research Trends, No. 26 http://www.researchtrends.com/issue26-january-2012/the-evolution-of-brain-drain-and-its-measurement-part-i/

5. Plume, A. (2012) "The evolution of brain drain and its measurement: Part II", Research Trends, No. 27 http://www.researchtrends.com/issue-27-march-2012/the-evolution-of-brain-drain-and-its-measurement-part-ii/

6. Small, H. & Klavans R. (2011). "Identifying Scientific Breakthroughs by Combining Co-citation Analysis and Citation Context". Paper presented at the Proceedings of 13th International Conference of the International Society for Scientometrics and Informetrics (ISSI 2011).