STI 2018 Conference Proceedings

Proceedings of the 23rd International Conference on Science and Technology Indicators

All papers published in this conference proceedings have been peer reviewed through a peer review process administered by the proceedings Editors. Reviews were conducted by expert referees to the professional and scientific standards expected of a conference proceedings.

Chair of the Conference

Paul Wouters

Scientific Editors

Rodrigo Costas
Thomas Franssen
Alfredo Yegros-Yegros

Layout

Andrea Reyes Elizondo
Suze van der Luijt-Jansen

The articles of this collection can be accessed at https://hdl.handle.net/1887/64521

ISBN: 978-90-9031204-0

23rd International Conference on Science and Technology Indicators (STI 2018)
"Science, Technology and Innovation indicators in transition"
12 - 14 September 2018 | Leiden, The Netherlands
#STI18LDN

Supplementing citations to PhD theses with citations from Google

Paul Donner*

*donner@dzhw.eu
Research Area "Research System and Science Dynamics", German Centre for Higher Education Research and Science Studies (DZHW), Schützenstraße 6a, Berlin, 10117 (Germany)

1 Introduction

1.1 The place of dissertations in scholarly communication and citation analysis

Completing a PhD by writing, defending and publishing a thesis is widely regarded as the standard entry into an academic career. The thesis usually contains the manifest outcomes of independent research. The thesis is formally assessed by graduation committee. Once exclusively published as monographs, the PhD thesis has been in a transition of form. The cumulative dissertation based on individually published scientific papers has become an important alternative to the traditional book in many disciplines.

Studies based on citation analysis have extended the knowledge of the role of doctoral theses and journal papers by PhD students. According to one study (Larivière, Zuccala and Archambault, 2008) dissertations only account for a very small fraction of cited references in the Web of Science database but the impact of individual theses was not investigated. The citation impact of journal articles to which PhD students contributed has only been studied on a large scale for the Canadian province of Quebec in Larivière (2012). The impact of journal papers with PhD student contribution is contrasted to all other papers with Quebec authors in the Web of Science database. As the impact of these papers, quantified as average of relative citations, is close to the comparison groups in three of four broad areas, we might tentatively assume that the impact is on par to non-PhD-student papers. The area with a notable difference between groups is arts and humanities, in which the coverage of publication output in the database is less comprehensive because a lot of research is published in monographs and in which presumably many papers will be written in French, another factor of lower coverage. No large scale study has been conducted on the impact of theses on the level of individual works or on the level of university departments. We so far have quite limited little information on the citation impact of theses.

1.2 Policy motivation

The education of future researchers is generally accepted as a core task of universities. Correspondingly, doctoral education indicators have been used in several university evaluation exercises, primarily in order to support committee peer evaluation judgments, for example in the Standard Evaluation Protocol (Netherlands) or the Forschungsrating pilot assessments (Germany). However, there is no consensus definition of a performance dimension of doctoral or young researcher education and support. It is not clear which

indicators provide a valid reflection of performance in such a dimension and whether they are redundant or complimentary, that is, if cover the same latent dimension or several separable aspects. To inform the development of reliable evaluation methods, it would thus be valuable to study the structure of relationships of doctoral education indicators on the institutional/departmental level. The citation impact of dissertations has, to the best of our knowledge, so far not been studied as a possible performance indicator on the level of departments. To test whether or not dissertation citation impact is a suitable indicator of doctoral education performance, citation data for theses needs to be collected, aggregated and studied for associations with other relevant indicators, such as doctorate conferrals, drop-out rates, graduate employability, thesis awards, or subjective program appraisals of graduates. As a first step towards a better understanding of doctoral education performance, we conducted a study on citation sources for dissertations. The present study is restricted to monograph form dissertations. However, to be able to assess the total scientific impact of a PhD project it is necessary to also include the impact of papers of cumulative dissertations and papers which are produced in the context of the PhD project which is formally only published in monograph form. For the time being, this is left to a later stage of the project.

1.3 Google Books as a complementary citation source

Google Books (GB) offers full text search of digitized and digital books. This can be used to find citing works by using the metadata of the target publications (cited works) as search terms. GB searches can be automated via an API. The search results need to be filtered to obtain valid citations. It has been shown that GB can be a valuable source of citation data in particular for books (Kousha & Thelwall, 2015). As can easily verified by the reader, the citation data which can be found in GB is not generally incorporated into Google Scholar.

1.4 Research objectives

The main research question addressed in this contribution is whether the collection and inclusion of GB citations can substantively supplement the citation data of monograph dissertations as obtained from Scopus and thereby improve the validity of citation analysis of these works with a perspective towards the study of relationships between thesis citation impact and other department-level doctoral education indicators in the future.

A secondary objective is to learn about the proportions of obtainable additional citations for different disciplines. Given that larger shares of publications in the social sciences and humanities, as opposed to science, technology, engineering and medicine disciplines, are in the form of monographs and written in languages other than English and that these types of publications are covered less completely in major citation databases, we may reasonably expect higher relative additional citation shares for the former discipline group to be extracted from GB. We are also interested in the relative values of citation rates of dissertations in Scopus and GB across disciplines.

2 Methods

2.1 Target data set of German theses (metadata)

The target data set of German dissertations for which citation data are sought, were obtained from the online catalog of the German National Library, sections H (publications of higher education institutions) and O (online publications), for the publication years 1996 to 2016. The data was carefully de-duplicated as one thesis may have several catalog entries for print, digital and published monograph versions. The full data set contains metadata for 421,526

theses. A large share of these are medical dissertations, which are required to obtain the degree of Dr. med., which a majority of medicine graduates aspiring to become practicing physicians do obtain. These theses are usually considered less substantial than regular dissertations, in fact, persons interested in medical research often obtain an additional Dr. rer. nat. degree subsequently or alternatively. It is not possible to distinguish these two types of medical dissertations in the data set, but together they make up about 193,000 theses.

2.2 Citation data collection

Scopus as a primary citation data source

In the most comprehensive study of theses' citation impact, Larivière, Zuccala and Archambault (2008) generated their set of thesis citations by searching for the term "thesis*" in the Web of Science cited reference indexes. There are some drawbacks to this method. Due to a lack of process documentations there is no way to be sure that the database producer consistently labels all cited theses, let alone is able to identify all cited theses in reference lists. Theses cited without noting their nature as a dissertation will be missed. Furthermore, up until very recently the "source" field of the cited reference index of WoS contained very short and inconsistently abbreviated title information. Because of this limitation, a search for known thesis titles was not feasible. Their method is therefore likely to miss some theses citations and this proportion cannot be estimated. A clear advantage of said method is the high precision which is achieved. The authors found very few citations to works which are not theses in their results.

A different approach was used in the present study. We decided on Scopus as a primary citation source because of its extensive coverage of the journal literature and its continuous increase in the coverage of books. But most importantly, its cited reference index contains full source title and document title information, if given in the original reference list. This makes it possible to directly search for known thesis titles.

To obtain citations to the theses in the data set from the references indexed in Scopus we proceeded as follows. According to its Content Coverage Guide (version of August 2017) Scopus does not index theses as primary documents, all cited theses must therefore be non-source documents. We therefore limit the set of candidate references by considering only non-source references, that is, references which are not linked to an item covered in Scopus. Among these, we accept a reference as a match to a target thesis if they have

- strictly identical author lastnames and first initial of given name (full given name is not available in Scopus if it is not stated in the original publication)
- publication year within ± 1 year of thesis publication year
- titles (in Scopus either source title or work title field) similar, such that when the longer title is truncated to the string length of the shorter title, these two strings have an edit distance similarity greater than 0.75.

These conditions were found to give qualitatively very good results in terms of precision and recall while being also computationally not too demanding. However, a comprehensive formal quantitative validation was not attempted.

Google Books citation data

Google Books citations were obtained with the Webometric Analyst tool by the Statistical Cybermetrics Research Group of University of Wolverhampton (Kousha & Thelwall, 2015).

The searches were performed in batch mode with the standard settings. The used search fields were author lastname, title (first six words), the name of the degree granting university's town, and publication year. The basic cleaning process implemented in the tool was used.

The obtained citation candidates were loaded into an SQL database for further cleaning steps and analysis. Lists of the most often occurring citing document titles, author names and publishers were created and the top results manually checked for non-academic sources. A number of institutional annual reports, topical bibliographies and historical institutional bibliographies were identified and all their citations removed (lists of 33 titles and 3 creator field values). Furthermore, the contents of the fields creator, title and publisher of the citing work returned from GB were found to be very messy and had to be cleaned.

2.3 Data combination

To remove duplicate citations resulting from the combination of the two citation sources, we deleted all GB citations which had identical title, author and publication year data to citations from Scopus. This resulted in approximately 1100 citations. Further, all GB citations which had a citing publication title which was identical to a journal title covered in Scopus were deleted (~800 citations). As the final step, GB and Scopus citation counts were calculated for all theses and loaded into one common database table for analysis. No restriction for a citation window was imposed.

3 Results

Theses are classified by subject by the German National Library based on its subject categories, which are based on a version of the Dewey Decimal Classification[1]. There was a change in the classification system in 2004. Based on the mapping between the two versions, a slightly aggregated custom version is used here, with approximate translations of the class labels in order to simplify the presentation and interpretation. The results are presented in table 1 for 41 subjects and in total.

We were able to identify about 140,000 citations to German theses in Scopus and a further 90,000 citations in GB, an increase of over 60%. The Pearson correlation of the citation counts of the two sources, computed as log(citations + 1), is 0.35. This figure is to be interpreted very cautiously, as the two distributions are highly skewed, discrete with a lot of ties at zero and have low averages (Thelwall, 2016). In fact, as pointed out by Thelwall (2016), there is a substantial correlation between the combined average citation count per field and the correlation coefficient of GB and Scopus citations per dissertations in the field (r=0.59). Taken at face value, Scopus and GB citation counts are moderately correlated. A dissertation cited at some relative level of the citation distribution in Scopus will be at a similar position in the citation distribution of GB, but with a substantial amount of variation.

[1] cf. http://www.dnb.de/EN/Erwerbung/Inhaltserschliessung/sachgruppenDnb.html

Table 1. Results by scientific field

Field	theses	ratio of citations in GB to those in Scopus ↓	average citations in GB	average citations in Scopus	total citations GB	total citations Scopus
Geosciences	6890	0.21	0.31	1.48	2156	10189
Chemistry	28147	0.22	0.05	0.23	1472	6545
Electrical engineering	5406	0.22	0.17	0.79	918	4246
Biology	33054	0.22	0.07	0.33	2414	10915
Chemical engineering	2438	0.22	0.23	1.03	550	2515
Physics, astronomy	24161	0.23	0.1	0.43	2366	10344
Veterinary medicine	11572	0.23	0.12	0.54	1441	6226
Mechanical engineering	13525	0.23	0.27	1.15	3638	15578
Mathematics, statistics	7050	0.26	0.39	1.49	2754	10499
Industrial engineering	4847	0.28	0.29	1.02	1405	4950
Computer science	8095	0.29	0.51	1.76	4156	14263
Technology general	1895	0.3	0.46	1.52	870	2873
Natural resources, energy, environment	2517	0.3	0.2	0.65	499	1640
Agriculture	1178	0.42	0.53	1.26	625	1480
Medicine, health	193199	0.53	0.03	0.07	6704	12593
Natural sciences general	474	0.66	0.27	0.4	127	191
Linguistics	657	0.76	0.72	0.94	474	620
General science and culture	295	0.78	0.63	0.81	186	238
Trade, communication, traffic	1605	0.92	0.68	0.74	1094	1194
Psychology	3350	1.19	0.54	0.45	1815	1522
Environmental protection and engineering	667	1.37	0.81	0.59	540	395
Library and information science,	331	1.39	0.8	0.57	265	190

archiving						
No subject	1105	1.4	0.27	0.2	303	216
Archeology and prehistory	620	1.44	1.23	0.85	760	526
English language and literature	961	1.46	0.54	0.37	520	357
Economics	19329	1.53	0.7	0.45	13441	8781
Architecture	1553	1.61	0.7	0.43	1083	671
Political science, military	5780	1.79	0.68	0.38	3908	2179
Ethnology	901	2.06	0.57	0.28	512	249
Sociology	4179	2.07	0.96	0.46	4008	1940
Languages and literatures other than English and German	1849	2.11	0.81	0.38	1497	709
Educational science	4366	2.33	0.75	0.32	3262	1399
Philosophy	1803	2.96	0.8	0.27	1442	487
Home economics, hospitality management studies	136	3.11	0.87	0.28	118	38
German language and literature	2691	3.14	0.92	0.29	2480	789
Religion	3464	3.31	1.31	0.39	4526	1366
Music and performance arts	1660	3.34	0.74	0.22	1223	366
Journalism	482	3.39	1.19	0.35	570	168
History	1859	3.89	1.76	0.45	3261	839
Visual arts	2082	4.21	0.89	0.21	1857	441
Law	15353	5.81	0.53	0.09	8209	1414
total	**421526**	**0.63**	**0.21**	**0.34**	**89449**	**142141**
total without Medicine, health	227223	0.64	0.36	0.57	82442	129332

Most dissertations remained uncited. In the Scopus data, 87% of theses are not cited; in the GB data, 90% are not cited. However, the combination of the two citation sources reduces the overall uncited rate to 82%. Theses are cited on average 0.55 times. Excluding the atypical field of medicine, the figure is 0.93, which supports the introductory remarks on medical dissertations.

In Scopus, the most often cited dissertations are in the fields of computer science, mathematics and statistics, and geosciences. In contrast, the highly cited fields in the GB citation data are history, religion, and archeology and prehistory. When combining the two citation sources we find that dissertations in chemistry, biology, physics and astronomy are relatively rarely cited. Theses in archeology and prehistory, history and computer science, on the other hand, are cited relatively frequently. The addition of GB citations leads to only small increases in citation average citation frequency in the natural and engineering sciences. Relatively large increases are observed in the social sciences and humanities.

Discussion

This study sought to quantify the amount of additional citations which can be obtained by supplementing citations data from a commercial bibliometric database with citations from Google Books for a type of rarely cited publications, namely PhD theses. We have argued that the retrieval of thesis citations in citation databases benefits from the presence of full cited work titles in the reference data and known thesis titles, which facilitates an analysis on the level of individual graduates or theses. We showed how the inclusion of GB citations results in overall very substantial numbers of additional citations, which would lead to increased validity of any further use of citation information for theses. The increase was particularly pronounced in the social sciences and humanities. Nevertheless, dissertations must still be considered rarely cited works.

References

Kousha, K., & Thelwall, M. (2015). An automatic method for extracting citations from Google Books. Journal of the Association for Information Science and Technology, 66(2), 309-320.

Larivière, V. (2012). On the shoulders of students? The contribution of PhD students to the advancement of knowledge. Scientometrics, 90(2), 463-481.

Larivière, V., Zuccala, A., & Archambault, É. (2008). The declining scientific impact of theses: Implications for electronic thesis and dissertation repositories and graduate studies. Scientometrics, 74(1), 109-121.

Thelwall, M. (2016). Interpreting correlations between citation counts and other indicators. Scientometrics, 108(1), 337-347.