On the effects of database errors on citation-based and on diversity indicators

Bart Thijs[1] and Mehmet Ali Abdulhayoglu[2]

[1] *bart.thijs@kuleuven.be*
KU Leuven, ECOOM, MSI, Naamsestraat 61, 3000 Leuven (Belgium)

[2] *mehmet.abdulhayoglu@kuleuven.be*
KU Leuven, ECOOM, Naamsestraat 61, 3000 Leuven (Belgium)

Abstract

This paper describes an ongoing research project that aims at capturing the extent of the problem with erroneous references in bibliographic databases. The study tackles three questions: What is the extent of the problem? What is the effect of these errors on bibliometric indicators and how these problems be solved. The first experiments with randomly added and removed references show low sensitivity of the studied citation-based indicators while substantial effect on disparity, one of the interdisciplinarity measures. Further research is needed to develop an appropriate procedure to detect and correct these errors.

Introduction

The calculation of bibliometric indicators being used for research evaluation or assessment must have the highest possible precision at any level since it is often an important input to decision-making processes such as funding allocation, promotion, or other science policy related requests. Consequently, building these indicators on reliable data is key to those conducting the analyses.

Recently, we detected errors in the indexed cited references in Web Of Science while validating scores and values in the framework of the measurement of interdisciplinarity. The measures we use (see Glänzel & Debackere, 2021) rely on the disciplines of the indexed cited references. These citation-based links between citing and cited document provide valuable information for other applications in bibliometrics like topic detection, document retrieval or similarity calculations. On the one hand, such erroneous records might not have significant effect on an analysis like topic detection assuming that the number of wrongly indexed references should be very few (a study about the effects of erroneous records on different type of analysis has been being carried out at our institute). On the other hand, its adverse effect is sometimes obvious as occurred to us when calculating IDRs (for individual assessments, policy makers might want to know how interdisciplinary the researchers' work are) of papers whose details are given in the subsequent section.

In this research in progress paper, we investigate the issues related to incorrectly indexed references in WoS and focus on three aspects:

1. How large is the problem of incorrect indexed cited references?
2. What is the effect of these errors on bibliometric indicators?
3. What tools and procedures are available to solve the problem?

Data Sources

Web Of Science

Clarivate's Web Of Science is ECOOM's main bibliographic database for scientometric studies. In this paper we use publications indexed in 2018. We calculate citation indicators using a three year citation window and take all indexed cited references into account for the measures related to interdisciplinarity.

Scopus

Elsevier's Scopus is our first go-to source to make the comparisons with WoS. Data is retrieved from Scopus using the API and from an inhouse copy. Both databases are linked using common identifiers like DOI or PubMedID. If these identifiers are lacking, standard bibliographic data like journal, author, title, publication year is used for record matching. Retrieval through the API is bound by certain restrictions.

Crossref

Crossref is one of the main free bibliometric data sources. Through its REST API, reference lists can be retrieved for given DOIs. On one hand, compared to WoS and Scopus, its API's rate limits are higher allowing a user to process large-scale data. On other hand, its servers can be quite busy hindering the user from a smooth process.

Openalex

As a new data source (some features are still in beta), it allows users to fetch data including cited references of papers through their REST API. Even though the rate limit is again an issue, it allows users to use "OR condition" within query that is, up to 50 DOIs can be searched with one query at a time. This makes it very handy to retrieve large scale data. However, in terms of data completeness, it lags behind the three sources mentioned so far. Since it is sort of a beta version, it can be a valuable source in the future.

Unpaywall

Another open database offering its data through an API requiring DOIs as input. It provides data for only open source papers and the bibliographic meta-data it offers is limited so no reference lists are provided. Nevertheless, it returns full text URL for PDF.

Results and discussion

Question 1. How large is the problem

Previous studies (eg Buchanan, 2006 or Van Eck & Waltman, 2016) investigated the nature and extent of errors in bibliographic database. Mapping errors were reported to range between 1.2 and 6.9 percent depending on the type of error.

During our manual inspection, we see that some authors separate references not one by one but according to their relevance. For example, they present sub-references as (a), (b)… or (i), (ii)… under one main reference. So, parsing becomes more challenging for the data providers leading wrong or inconsistent indexes. Or in some papers we come across, there are more than one references section which for sure confuses parsing logics of different data providers resulting in different reference lists for the same paper in different sources.

When further investigating the extend of the problem, we departed from the calculation of disparity and variety scores (see below) where we found for a limited set of papers an extremely high disparity score. Such scores can be explained by the presence of one or a few references to publications assigned to fields quite distant to the fields or disciplines of most references, (eg. a reference to poetry included in a paper on high energy physics). We investigated 100 of the top ranked papers with respect to disparity and concluded that 37 of these have at least one incorrect reference. Some of the references were incorrectly indexed in the database due to text processing or parsing errors. Other references were linked to the incorrect target paper due to combinations of identical bibliographic data like publication year, volume, first author name, or first page.

Question 2. What is the effect on bibliometric indicators.

We try to answer this question by looking at the changes in bibliometric indicators when introducing random errors in the underlying citation data. In a first step, we used the thresholds calculated for the classification of citation distributions in the framework of the CSS-methodology and analogously for disparity and variety. The CSS-methodology calculates – in an iterative process- the average number of citations (or disparity and variety) in truncated distributions (see for more details).

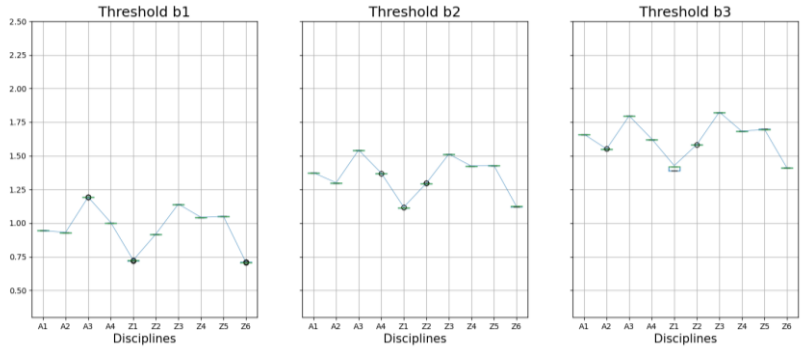

Figure 1. Boxplot and line for the sampled and the original values of the citation-based thresholds for disciplines in Agriculture (A) and Biology (Z).

Figure 1. presents the results for citation distribution. In a random sampling procedure, citations are added or removed from target publications. 5% of the publications indexed in 2018 are affected. Half of those affected get one additional citation while the other half loses a citation. This procedure results in a situation that corresponds with the error of linking a reference with the incorrect target publication. Such erroneously linking results in the incorrect increase of the number of citations of one paper and the decrease on another. This random procedure is repeated 40 times and each run results in a different set of thresholds for the CSS-methodology. These sampled values are presented in the boxplots in figure 1 while the line indicates the original value for each of the thresholds across the disciplines. It can be observed in figure 1 that the random introduction of errors in 5% of the publications in Agriculture and Biology has no effect on the calculations of the thresholds. There is hardly any variance in the 40 random values and they do not deviate for the original score. This pattern can be seen for all disciplines. Furthermore, the distribution of publications over the four citation classes is for each of the 40 random runs very close to the original distribution.

For the analysis for the effect of erroneous cited references on the interdisciplinarity measures disparity and variety we use a much lower error rate of 0.5%. For half of the affected publications, we added a random reference with associated discipline to the cited references, in the other half, one random reference was removed. This procedure preserves the total number of total number of references. Figure 2 plots the thresholds for disparity. This measure is calculated following a rewrite of the Leinster-Cobbold disparity and reads as:

$$ {}^{2}D^{S} = \left(\sum_{i,j=1}^{N}(1 - d_{ij})p_i p_j\right)^{-1}, \quad (1) $$

where

- N is the number of classes or topics in the applied classification system,
- p_i denotes the proportion of class i in the total set, and
- d_{ij} is the dissimilarity between classes i and j or expressed as similarity s_{ij} taking a value

We refer to the Zhang et al (2016) and Glänzel and Debackere (2021) for more details on the measures of disparity and variety

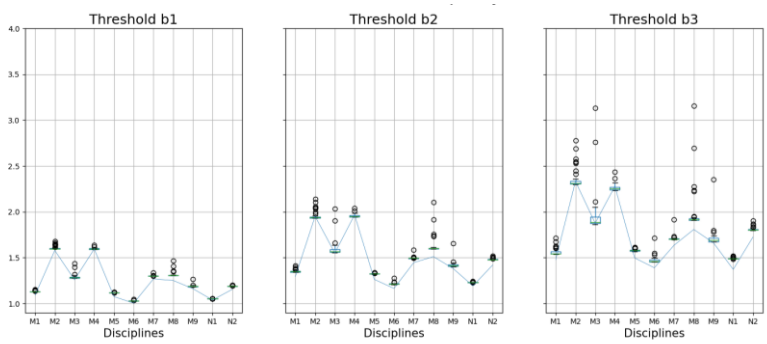

Figure 2. Boxplot and line for the sampled and the original values of the disparity-based thresholds for disciplines in Non-Internal Medicine (M) and Neurology and Behavioral Sciences (N).

As opposed to figure 1, we observe much more variance in the values obtained from the sampling procedure. Almost all values are above the original value. In fact, it seems that removing cited references, and thus probably lowering the disparity, has little or no effect on the thresholds while adding random references, with distant disciplines adds substantial to the resulting values. Especially for the third threshold, the variance is substantial. This means that the boundaries of the classes denoted as high or very high disparity are very sensitive for errors in certain fields. Other fields show similar patterns.
The same analysis is also repeated for variety and is presented in figure 3. Here we use Inverse Simpson Index (Glänzel & Debackere, 2021, p13)

$$^2D = (\sum_{i=1}^{N} p_i^2)^{-1} \quad (2)$$

Here we see sampled scores that are a bit lower than the original values. But, given the fact that the thresholds for this measure are quite close to each other, more research is needed to investigate the effect on the classification of papers in the four classes.

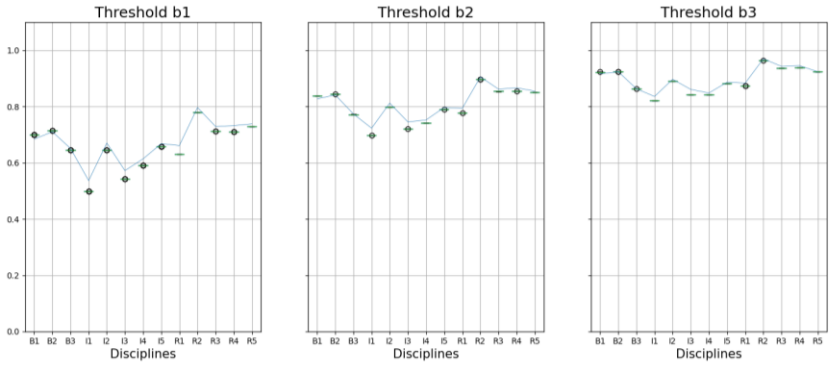

Figure 3. Boxplot and line for the sampled and the original values of the variety-based thresholds for disciplines in Internal Medicine (I) Biosciences (B) and Biomedical Research (R).

Question 3. How can we solve this problem

Here we list some simple ideas/steps to deal with solving the issue. The aim is not to provide a fully automated approach but manage the process at least partly automated so that the effort for manual checks remains reasonable. At this point, the approach is still quite crude.

1. Comparing the number of references for each paper in WoS and Scopus is the first step. Both Scopus and WOS have online API's that enable document searches based on identifiers like DOI.
2. Since an equal number does not guarantee that the references are correctly indexed, title pairs are constructed from the two sources and their Levenshtein similarity scores based on character based n-grams are calculated and the pairs having a similarity score lower than 0.85 are filtered out. Aside from title matching, some other matching rules such as identical first page, journal, volume, issue, author etc. will be added for the most reliable matching results. If all the WoS references are matched after the filtering, we can be sure that references are correct under the assumption that both sources are not incomplete. From our sample data (100 papers with the highest IDRs), the paper with the highest IDR score, we spot that 2 papers are wrongly indexed in WoS while all are correct in Scopus.
3. If WoS and Scopus are not in line regarding the number of references or Scopus does not index the WoS paper in question, Crossref could be beneficial if it has the corresponding reference data. Then, the steps from the previous bullet could be applied. From our sample dataset for some of which Scopus fails to return complete reference lists, Crossref seemed promising as it could offer a complete reference set. However, again as is the case for WoS and Scopus, Crossref often returns incomplete reference lists.
4. So, the real challenge is when those three sources (WoS, Scopus, Crossref) have different number of references. Then, the use of the original full text seems to be the best option. Tools like Grobid[1] aim at parsing and restructuring scientific papers and obtain pretty good results (F1-scores of 0.89) but often fail for those publications where also the main bibliographic database are not successful.

Conclusion

Previous research complemented with our own investigation reveal that the problem with incorrect references and erroneous links to target papers is quite substantial in the main bibliographic databases. But, it is reassuring that the effect on citation-based indicators can be considered to be marginal as showed by the repeated introduction of random errors. With respect to the disparity score, the user has to be cautious as this indicator is much more sensitive to these errors. Especially the introduction of additional references and cited disciplines can distort the indicators. For the other measure related to interdisciplinarity, the results are not unambiguously. The sampled values are only slightly lower but the effect on the distribution among classes is not clear.

References

Buchanan, R. A. (2006). Accuracy of cited references: The role of citation databases. College & Research Libraries, 67(4), 292-303.

Glänzel, W., & Debackere, K. (2021), Various aspects of interdisciplinarity in research and how to quantify and measure those. *Scientometrics*, 10.1007/s11192-021-04133-4

Van Eck, N.J. & Waltman, L. (2017) Accuracy of citation data in Web of Science and Scopus. *Proceedings of the International Conference on Scientometrics and Informetrics*,1087-1092

Zhang, L., Rousseau, R., & Glänzel, W. (2016). Diversity of references as an indicator of the interdisciplinarity of journals: taking similarity between subject fields into account. *Journal of the association for information science and technology*, 67(5), 1257-1265.

[1] See https://grobid.readthedocs.io/en/latest/