Journal of Informetrics 10 (2016) 933–953

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Empirical analysis and classification of database errors in Scopus and Web of Science

Fiorenzo Franceschini*, Domenico Maisano, Luca Mastrogiacomo

Politecnico di Torino, DIGEP (Department of Management and Production Engineering), Corso Duca degli Abruzzi 24, 10129, Torino, Italy

ARTICLE INFO

Article history:
Received 9 March 2016
Received in revised form 15 June 2016
Accepted 6 July 2016

Keywords:
Data accuracy
Database error
Omitted citation
Error classification
Phantom citation
Scopus
Web of Science

ABSTRACT

In the last decade, a growing number of studies focused on the qualitative/quantitative analysis of bibliometric-database errors. Most of these studies relied on the identification and (manual) examination of relatively limited samples of errors.

Using an automated procedure, we collected a large *corpus* of more than 10,000 errors in the two multidisciplinary databases Scopus and Web of Science (WoS), mainly including articles in the Engineering-Manufacturing field. Based on the manual examination of a portion (of about 10%) of these errors, this paper provides a preliminary analysis and classification, identifying similarities and differences between Scopus and WoS.

The analysis reveals interesting results, such as: (i) although Scopus seems more accurate than WoS, it tends to forget to index more papers, causing the loss of the relevant citations given/obtained, (ii) both databases have relatively serious problems in managing the so-called *Online-First* articles, and (iii) lack of correlation between databases, regarding the distribution of the errors in several error categories.

The description is supported by practical examples concerning a variety of errors in the Scopus and WoS databases.

1. Introduction

Bibliometric databases are commonly adopted by individual scientists and research institutions for (i) searching scientific documents, (ii) providing information on the citation impact of the scientific output, and (iii) supporting the selection of the scientific journals were to publish.

The abundance of bibliometric and/or bibliographic disciplinary databases (e.g., PubMed, MathSciNet, PsycINFO, IEEEXplore, EconLit, etc.) contrasts with the relatively limited number of multidisciplinary databases: Google Scholar (GS), Scopus, and Web of Science (WoS). A peculiarity of GS is to automatically index publications/citations through web crawlers, which allows to achieve considerably more coverage than Scopus and WoS. In fact, GS is estimated to contain approximately 160 M total documents, while Scopus approximately 13 M and WoS approximately 10 M (Mongeon & Paul-Hus, 2016; Orduna-Malea, Ayllón, Martín-Martín, & López-Cózar, 2015). Unfortunately, the automatic indexing of GS inevitably causes many errors (Labbé, 2010) and (almost) completely disqualifies GS with respect to its two competitors, to the extent that most consider GS simply as a search engine, certainly not a serious bibliometric database. Nevertheless, some recent studies indicate that the GS data quality is gradually improving (Moed, Bar-Ilan, & Halevi, 2016; Prins, Costas, van Leeuwen, & Wouters,

* Corresponding author.
E-mail addresses: fiorenzo.franceschini@polito.it (F. Franceschini), domenico.maisano@polito.it (D. Maisano), luca.mastrogiacomo@polito.it (L. Mastrogiacomo).

http://dx.doi.org/10.1016/j.joi.2016.07.003
1751-1577/

2016). Furthermore, the data quality of GS, Scopus and WoS were discussed in a number of comparative studies addressing coverage and overlap (e.g., Archambault, Campbell, Gingras, & Larivière, 2009; Harzing & Alakangas, 2016; Meho & Yang, 2007; Mikki, 2010; Wang & Waltman, 2016; Wildgaard, 2015).

In the last two years, we have been investigating the Scopus and WoS errors, analysing the so-called *omitted citations* – i.e., missing links between citing and cited papers – which represent one of the major consequences of database errors (Franceschini, Maisano, & Mastrogiacomo, 2013). An interesting result – which corroborates the findings of previous studies (Buchanan, 2006; Hildebrandt & Larsen, 2008; Larsen, Hytteballe Ibanez, & Bolling, 2007; Moed, 2002; Moed, 2005; Moed & Vriens, 1989; Olensky, 2015; Tunger, Haustein, Ruppert, Luca, & Unterhalt, 2010) – is that the omitted-citation rate of the two databases is far from being negligible: more than 4% for Scopus and more than 6% for WoS (Franceschini, Maisano, & Mastrogiacomo, 2014). We showed that the editorial style of some publishers can favour database errors and – although Scopus and WoS tend to be more and more careful in indexing new papers – they do little to correct the errors already present in the database (Franceschini et al., 2014; Franceschini, Maisano, & Mastrogiacomo, 2016a). Also, we came across many weird errors, discussed in a recent "opinion" paper (Franceschini, Maisano, & Mastrogiacomo, 2016b).

The majority of our past researches relied on the analysis of a relatively large *corpus* of scientific articles, consisting of almost 24,000 cited articles – confined to the Engineering-Manufacturing field – and almost 100,000 corresponding citing articles. Among these articles, thousands of omitted citations were identified using an automated algorithm, which requires the combined use of Scopus and WoS and is based upon the idea that the mismatch between the citations occurring in one database and another one is evidence of possible errors/omissions (Franceschini et al., 2013).

In our previous researches (Franceschini et al., 2013, 2014, 2016a; Franceschini, Maisano, & Mastrogiacomo, 2015a), we analyzed the Scopus and WoS omitted citations, studying the influence of several factors, such as journal or publisher of cited papers, issue year of citing papers, date of database queries, etc. However, we did not investigate the causes of these omitted citations – i.e., the nature of database errors – in a detailed and structured way.

Consistently with the categorization suggested by Buchanan (2006), (at least two) types of database errors can be defined:

A *Pre-existing errors*: errors made by authors/editors/publishers when preparing the list of cited articles for their publication; e.g., errors in the author name(s), article title, issue year, volume number, pagination, etc.
B *Database mapping errors*: failures to establish an electronic link between a cited article and the corresponding citing articles that can be attributed to data-entry errors in the database; e.g., transcription errors, cited article omitted from a cited-article list, etc.

While the errors in the first category are (at least partly) justifiable, being caused by inaccuracies in the original papers, those in the second one are introduced by databases, in the data-entry process.

The goal of this paper is to delve into the large corpus of omitted citations available from our past research and perform a statistical analysis of the relevant database errors, trying to answer to the following research questions:

- *What are the more frequent errors of Scopus and WoS and the similarities and differences between the two databases?*
- *Are the results of this research in line with those of other researches in the field of bibliometric-database errors?*
- *Does this research provide a representative picture of the Scopus and WoS errors?*
- *In the light of the results obtained, what are the practical implications to users and administrators of the Scopus and WoS databases?*

The proposed statistical analysis requires a thorough manual examination of the database records and the original cited/citing papers, with special attention to the cited-article lists. Due to the relatively large time consumption of this process, it will be limited to the 10% of the (more than 10,000) omitted citations available.

The remainder of the paper is organized into five sections. Section 2 recalls the automated algorithm for detecting omitted citations. Section 3 illustrates the analysis methodology in detail and presents some indicators for estimating the rate of the so-called *phantom-citations* of the two databases. Section 4 describes the analysis results; the description is supported by practical examples concerning various errors in Scopus and WoS. Section 5 summarizes the original contributions of this paper, describing its implications and limitations. Additional information is contained in the Appendix A.

2. Automated algorithm for analysing the omitted citations

Before recalling the algorithm, we present an introductory example to illustrate how it works. Let us consider a fictitious paper of interest, indexed by Scopus and WoS. The number of citations received by this paper is four in Scopus and six in WoS (see Table 1).

The union of the citations recorded by the two databases is a total of eight citations. Among these citations, only five come from sources (i.e., journals or conference proceedings) officially covered by both databases (highlighted in grey in Table 1). Focusing on these five *theoretically overlapping* (TO) citations, two are omitted by Scopus (but not by WoS) and one is omitted by WoS (but not by Scopus). Therefore, from the perspective of the paper of interest, a rough estimate of the omitted-citation rate is $2/5 \approx 40\%$ in Scopus and $1/5 \approx 20\%$ in WoS. The same reasoning can be extended to multiple papers of interest and more than two bibliometric databases.