Journal of Informetrics 12 (2018) 436–447

Contents lists available at ScienceDirect

Journal of Informetrics

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

Algorithmically generated subject categories based on citation relations: An empirical micro study using papers on overall water splitting☆

Robin Haunschild[a,*], Hermann Schier[a,1], Werner Marx[a], Lutz Bornmann[b]

[a] *Max Planck Institute for Solid State Research, Heisenbergstr. 1, 70569 Stuttgart, Germany*
[b] *Division for Science and Innovation Studies, Administrative Headquarters of the Max Planck Society, Hofgartenstr. 8, 80539 Munich, Germany*

ARTICLE INFO

Article history:
Received 7 August 2017
Received in revised form 22 February 2018
Accepted 14 March 2018

Keywords:
Field categorization
Bibliometrics
Algorithmically constructed classification systems
Subject categories

ABSTRACT

One important reason for the use of field categorization in bibliometrics is the necessity to make citation impact of papers published in different scientific fields comparable with each other. Raw citations are normalized by using field-categorization schemes to achieve comparable citation scores. There are different approaches to field categorization available. They can be broadly classified as intellectual and algorithmic approaches. A paper-based algorithmically constructed classification system (ACCS) was proposed which is based on citation relations. Using a few ACCS field-specific clusters, we investigate the discriminatory power of the ACCS. The micro study focusses on the topic 'overall water splitting' and related topics. The first part of the study investigates intellectually whether the ACCS is able to identify papers on overall water splitting reliably and validly. Next, we compare the ACCS with (1) a paper-based intellectual (INSPEC) classification and (2) a journal-based intellectual classification (Web of Science, WoS, subject categories). In the last part of our case study, we compare the average number of citations in selected ACCS clusters (on overall water splitting and related topics) with the average citation count of publications in WoS subject categories related to these clusters. The results of this micro study question the discriminatory power of the ACCS. We recommend larger follow-up studies on broad datasets.

1. Introduction

In bibliometrics, it is often necessary to compare the impact of publications from different fields.[2] However, it should be avoided to use bare citation counts ("times cited") from Web of Science (WoS, Clarivate Analytics) or Scopus (Elsevier) for such comparisons. Many bibliometric studies have shown that there are large differences in citation rates between fields, which cannot be explained by the quality of publications (see, e.g., Bornmann & Marx, 2015). Field-normalized indicators have been developed in bibliometrics which make cross-field comparisons possible. "The idea of these indicators is to correct

☆ "The peer review process of this paper was handled by Staša Milojević, Associate Editor of Journal of Informetrics".

* Corresponding author.
E-mail addresses: R.Haunschild@fkf.mpg.de (R. Haunschild), bornmann@gv.mpg.de (L. Bornmann).

[1] Hermann Schier passed away before the final version of the manuscript has been written. He gathered and analyzed many of the results and contributed to an earlier draft of this manuscript.

[2] We use the terms "field" and "topic" interchangeably, because the distinction between the two is rather arbitrary (Sugimoto & Weingart, 2015).

https://doi.org/10.1016/j.joi.2018.03.004

as much as possible for the effect of variables that one does not want to influence the outcomes of a citation analysis, such as the field . . . of a publication" (Waltman, 2016, p. 375). The use of normalized indicators in research evaluation is one of the guiding principles for research evaluation in the Leiden manifesto for research metrics (Hicks, Wouters, Waltman, de Rijcke, & Rafols, 2015).

In recent years, several methods have been proposed for the calculation of normalized citation scores. An overview of these methods can be found, for example, in Mingers and Leydesdorff (2015), Waltman (2016), and Bornmann and Marx (2015). Today, indicators based on the idea of counting highly cited publications are seen as a robust method for measuring citation impact across fields (Wilsdon et al., 2015). An important topic in the calculation of field-normalized indicators is the way in which research fields are defined, i.e. which field-categorization schema is used in bibliometrics to calculate the expected number of citations (Wilsdon et al., 2015).

The most common approach in bibliometrics is to work with subject categories defined by Clarivate Analytics in WoS or by Elsevier in Scopus. These subject categories are based on sets of journals publishing research from similar areas. However, the use of journal sets for field-normalization is heavily criticized. The most critical point is papers published in multi-disciplinary journals which cannot be assigned to corresponding fields using journal sets (Hui, 2015; Kronman, Gunnarsson, & Karlsson, 2010). Alternative approaches which can be used instead of journal sets have been classified by Wang and Waltman (2016) in mono-disciplinary and multi-disciplinary classification systems.

A mono-disciplinary classification system "covers publications in one particular research area and usually provides a classification at a relatively high level of detail" (Wang & Waltman, 2016, p. 348). Mono-disciplinary classification systems, as the Physics and Astronomy Classification Scheme (PACS, see https://publishing.aip.org/publishing/pacs/pacs-2010-regular-edition) system used in this study, are mostly expert-based approaches (Wang & Waltman, 2016) where experts in the fields (at least the authors of a paper) assign papers to corresponding subject categories. Nowadays, paper classification of PACS is supported by machine-indexing but expert controlled. PACS is included in the broader classification scheme of the INSPEC database. At the highest hierarchical level, INSPEC features the sections A (Physics), B (Electrical Engineering & Electronics), C, (Computers & Control), and D (Information Technology) (The Institution of Electrical Engineers, 1992). The section A of INSPEC is identical with PACS.

Waltman and van Eck (2012) introduced a method for algorithmically constructing classification systems (ACCS) at the level of individual publications. The method is a multi-disciplinary classification system and is based on citation relations between publications. The approach which is explained in more detail in Section 2 plays a prominent role among the available schemes, because it is used in the Leiden ranking (a university ranking based on bibliometrics, available at http://www.leidenranking.com/) for the calculation of field-normalized impact scores (the Leiden ranking uses a different solution of the ACCS than the system studied here). The method employed by Waltman and van Eck (2012) uses direct citation relations between papers for classification. They provide software referred to as "Modularity Optimizer" which uses direct citations as similarity measure. The general ACCS concept can be used, however, with different similarity measures (e.g., bibliographic coupling, co-citations, and textual comparison).[3] The results of Klavans and Boyack (2017) show in general that classification systems based on journal schemes are inaccurate compared to algorithmically constructed classifications. Sjögårde and Ahlgren (2018) discuss "how the resolution parameter given to the Modularity Optimizer software can be calibrated so that obtained publication classes correspond to the size of topics" (p. 149).

In this case study, we investigate the ability of the method to reliably assign publications to fields. This study is not intended to undertake a broad comparison between ACCS and other field classification systems, but to analyze one specific field, namely, "overall water splitting", in more detail. The use of this field has three advantages: (1) Most of the publications can be reliably compiled in WoS by a topic search. (2) One of the labels for a cluster in ACCS is "overall water splitting" (cluster 3.7.3). (3) Three of the four authors of this paper have a background in chemistry and physics. Therefore, we are able to provide a qualitative perspective on the search results. Research on overall water splitting is important for hydrogen gas production from water. The direct water splitting using solar cells or other renewable energy sources is especially appealing. Such a detailed and qualitative approach is not possible on a large scale, like the study by Klavans and Boyack (2017).

In the empirical part of this study, we present the results of several analysis steps: (1) Experts examined (read) a sample of papers in ACCS cluster 3.7.3 to determine whether they really deal with the topic "overall water splitting". (2) We investigate the spread of publications found by the WoS topic search over the ACCS clusters: Are most of the "overall water splitting" publications assigned to cluster 3.7.3? (3) We take the other way around and study the spread of publications in the ACCS cluster "overall water splitting" over WoS and PACS subject categories (SCs). (4) We compare the ACCS cluster 3.7.3 with related clusters of similar size (3.7.2 and 3.7.4) to investigate the discriminatory power of the ACCS. Papers assigned to different clusters should differ in terms of content. (5) We study citation impact differences of the papers in these related clusters of similar size. The clusters on the same hierarchical level are ordered by the number of papers in the cluster.

2. Field classification systems used in this study

Science is structured by disciplines (e.g. physics or chemistry), whereby each discipline is a specific domain of particular research traditions including paradigms, codes of practice, and methods (Ziman, 1996). Although it is practically impossible

[3] For the sake of brevity, we refer to the classification system studied here as ACCS.