Journal of Informetrics 7 (2013) 63–71

Contents lists available at SciVerse ScienceDirect

Journal of Informetrics

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

Investigating the universal distributions of normalized indicators and developing field-independent index

Jiang Wu

School of Information Management, Wuhan University, Wuhan, Hubei 430072, China

ARTICLE INFO

Article history:
Received 24 July 2012
Received in revised form 24 August 2012
Accepted 27 August 2012

Keywords:
g-Index
h-Index
Field independence
Universality
Power-law distributions
Lognormal distributions

ABSTRACT

Using the dataset based on Thomson Reuters Scientific "Web of Science" the distributions of some well-known indicators, such as h-index and g-index, were investigated, and different citation behaviors across different scientific fields resulting from their field dependences were found. To develop a field-independent index, two scaling methods, based on average citation of subject category and journal, were used to normalize the citation received by each paper of a certain author. The distributions of the generalized h-indices in different fields were found to follow a lognormal function with mean and standard deviation of approximately −0.8 and 0.8, respectively. A field-independent index *fi-index* was then proposed, and its distribution was found to satisfy a universal power-law function with scaling exponent α approaching 3.0. Both the power-law and the lognormal universality of the distributions verified the field independence of these indicators. However, deciding which of the scaling methods is the better one is necessary for the validation of the field-independent index.

1. Introduction

Science has become more multi-disciplinary, and the boundary of scientific field appears to be more undefined (Shiffrin, 2004). Thus, performances of scientists from different fields sometimes have to be compared using measurements (Ball, 2008). Currently, the h-index (Hirsch, 2005) and majority of citation-based impact measures are field dependent (Egghe, 2010). The development of field-independent index has attracted a number of attentions recently (Batista, Campiteli, & Konouchi, 2006; Iglesias & Pecharromán, 2007; Qiu, Ma, & Cheng, 2008). Ball (2008) pointed out that comparing achievements made by scientists from different fields is helpful in recruiting for a faculty position in case two candidates from different fields are being considered, and also beneficial in evaluating team performances of different scientific departments. Therefore, developing a field-independent index is necessary.

To develop a field-independent index, scholars usually argue the issues of scaling methods (Abramo, Cicero, & D'Angelo, 2012; Alonso, Cabrerizo, Herrera-Viedma, & Herrera, 2009; Waltman, Eck, Leeuwen, Visser, & Raan, 2011a; Waltman, Eck, Leeuwen, Visser, & Raan, 2011b). Iglesias and Pecharromán (2007) proposed the multiplication of the h-index of one author by the ratio of the average number of citations received by all papers published in the field of "Physics" to that in the field where the author belongs. They scaled the h-index directly instead of scaling each paper of that particular author to develop a field-independent index. Similarly, Batista et al. (2006) tried to scale the h-index using the total number of authors in the h papers as a relative field-independent index, and Raan (2006); Raan, Leeuwen, Visser, Eck, and Waltman, 2010 used a group-based average citation to normalize the citations of the papers published by an author. By contrast however, Radicchi, Fortunato, and Castellano (2008) and Radicchi, Fortunato, Markines, and Vespignani (2009) and Bornmann et al. (2009) scaled the citations of

E-mail address: jiangwu.john@gmail.com

1751-1577/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.joi.2012.08.007

each paper using an average number of citations in a subject category. Leydesdorff and Opthof (2010) also agreed that the citation of each paper would be more consistent if scaled separately to develop an index across different disciplines. They argued that each paper's citation should also be scaled using the average number of citations from each journal because the subject categories defined in the Web of Science (WoS) sometimes have difficulty in identifying to what field a paper belongs. Futhermore, cases exist where a paper belongs to more than one subject category. Although harmonic (rather than arithmetic) average has been proposed to handle these overlaping subject categories, justifying the scientific field remains difficult (Waltman et al., 2011a, 2011b). Therefore, scaling of a paper in terms of subject categories and journals should be understood further.

Using the WoS-based dataset, Radicchi et al. (2008) scaled the number of bare citations c of an article as $c_f = c/c_0$, where c_0 is the average number of citations received in a year by all the articles in that discipline published in the same year. They found that the distribution of the scaled number of citation c_f of articles can be universally fitted by a lognormal curve with a variance of 1.3. In addition, they also proposed a method of calculating a generalized h-index (*ghindex*) based on c_f and a new decreasing rank scaled by average number N_0 of articles published by all authors in a discipline (scientific field). Thereafter, Bornmann et al. (2009) validated c_f at the micro level using data from chemistry, and the proposed z-scores are better suited than c_f to produce universality of discipline-specific citation distributions. However, from their report (Bornmann et al., 2009), their finding was based on the chemistry journal Angewandte Chemie-International Edition, which needs to be tested at the macro level using extensive database. In reality, calculating c_0 is very time consuming because, until now, all database services, such as WoS, have not provided the value of c_0 for all subject categories over different years. Although the Aggregate Impact Factor of each subject category are provided in WoS, their calculations only depend on the citations received by the papers in the last two years and cannot be used as evidence to measure the cumulative impact of authors who have published in a certain subject category for a long period. Considering the lack of available statistical results in current databases, the WoS dataset was directly utilized to calculate c_0 and to investigate the distributions of scaled indicators across different scientific fields to evaluate the scientists.

In this paper, therefore, empirical investigation of the papers and their authors is done based on WoS from 1980 to 2009. This paper aims to investigate the distributions of the indicators used to measure the impact of an author. The cumulative citations of each paper in a year are scaled based on the journal-based and subject-category-based scaling denominators in the same year. Both original and scaled citations are applied to calculate the indicators for comparison purposes. Furthermore, a field index is proposed to measure the success of an author in a specific field (i.e., subject category or journal), and all the field indices where the author has been involved are integrated to propose a field independent index *fi-index* to measure the career success of an author. The h-indices, g-indices, mean citations, and total number of publications of an author from six scientific fields are first investigated for comparison to show their field-dependent characteristics, i.e., different citation and publication behaviors in different scientific fields can be observed. Egghe (2007) applied aspects of linear three-dimensional Lotkaian informetrics to determine the distributions of h-index and g-index, and this paper further empirically investigates these distribution using the WoS dataset. Then, the distributions of the generalized *h-index* (*ghindex*) and the proposed *fi-index* are also investigated to find the universality of these distributions to verify the field-independent characteristics of the *ghindices* and *fi-indices*. To examine their independences of scientific fields, this paper utilizes the investigation of the universal distribution of an indicator with a universal fitted curve, which was commonly used for testing field independence (Petersen, Wang, & Stanley, 2010; Radicchi et al., 2008). The proper choice of the journal-based or subject-category-based scaling method on developing a field-independent index is also studied in this paper.

This paper is organized as follows: in Section 2, the dataset and the methods employed to investigate distributions and to calculate the field-independent index are reviewed. In Section 3, the results and universality of distributions are presented to shed light on developing the field-independent index. In Section 4, conclusions are drawn and possible extensions of this work are presented.

2. Data and methods

Dataset based on Thomson Reuters Scientific "WoS", spanning from 1980 to 2009 was adopted. The scientific field is extracted from the first part of each subject category (the part before the comma), e.g., for subject category "Computer Science, Cybernetics," "Computer Science" is extracted as the name of this field. Therefore, each field can include several subject categories in WoS. If a paper in WoS belongs to more than one subject categories, all the subject categories of this paper will be extracted accordingly. To investigate the distribution of the indicators in the different fields, six scientific fields, namely, "Chemistry," "Computer Science," "Information Science," "Oncology," "Nanoscience," and "Economics" are selected. Active authors who have at least one publication in their correspoding field from 2008 to 2009 are extracted to compare their impact in terms of different indicators based on their papers from 1980 to 2009. The Oncology and Nanoscience fields are selected because authors usually focus on this narrow field, and these fields could possibly include all the papers written by a particular author. The Chemistry, Computer Science, Information Science, and Economics fields are selected because they present greater possibility of showing different publication and citation behaviors in this wide field that belongs to natural or social science. In WoS, all papers classified as "article," "letter," and "review" types in the corresponding fields are extracted for further analysis. The description of the dataset used in this paper is shown in Table 1.

To caculate indicators of authors in the underlying dataset, author name disambiguation is carefully considered in this paper. Author name disambiguation remains an open problem (Smalheiser & Torvik, 2009). The biggest challenge in the WoS dataset is that the same initial and last names may be used by many different individuals (e.g., some common names may