Journal of Informetrics 6 (2012) 689–699

Contents lists available at SciVerse ScienceDirect

Journal of Informetrics

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

An extension of the *h* index that covers the tail and the top of the citation curve and allows ranking researchers with similar *h*

Miguel A. García-Pérez*

Universidad Complutense, Madrid, Spain

ARTICLE INFO

Article history:
Received 13 April 2012
Received in revised form 8 June 2012
Accepted 10 July 2012

Keywords:
Citation analysis
Hirsch index
h index
Scientometric indicators

ABSTRACT

Citation curves for researchers with the same *h* index can vary greatly in the heaviness of their top (excess citations to core papers) or the heaviness of their tail (citations to non-core papers), revealing quantitative differences across researchers. Also, promotion to the next higher *h* depends only on citations received by a small subset of papers, so that researchers with a given *h* may have citation curves whose top and tail reveal a weaker impact than that of researchers with a lower *h*. To overcome these problems, we propose a *two-sided h index*, an extension that computes additional *h* indices progressively up the top and out the tail of the citation curve. This extension represents a *citation curve descriptor* one of whose elements is the scalar *h*. The advantages of the two-sided *h* index are illustrated through analysis of citation curves for 88 researchers with *h* indices ranging from 8 to 20. Several schemes are also discussed that use the two-sided *h* index to define criteria for ranking researchers within and across scalar *h* indices, according to whether the top of the citation curve, its tail, or both are deemed relevant under the circumstances in which research accomplishments are assessed.

1. Introduction

The *h* index (Hirsch, 2005) soon gained popularity and influence as a measure of the impact of an individual's research output (Zhang, Thijs, & Glänzel, 2011). The *h* index is the largest number of an individual's publications that have received at least *h* citations each. These *h* publications are referred to as the *h* core. The strengths and weaknesses of the *h* index have been thoroughly investigated and numerous variants and extensions have been proposed. Panaretos and Malesios (2009) and Rosenberg (2011) present detailed reviews and analyses of these variants and extensions, which vary from alternative definitions of the core to the use of schemes that incorporate non-core publications or that compensate for the effects of multiple authorship, scientific age, or self-citations. Bornmann, Mutz, Hug, and Daniel (2011) and Schreiber, Malesios, and Psarakis (2012) presented large-scale studies showing that many of these alternative indices are highly correlated with the *h* index, which suggests that they provide redundant information.

Computation of the *h* index faces two practical problems, even when it is done by the researchers themselves and, thus, with full knowledge of the set of papers for which citation counts must be retrieved. Many commercial and open-access databases are available that provide citation counts but their sources differ and, thus, they return meaningfully different numbers of citations for the same paper (Bar-Ilan, 2008; García-Pérez, 2010; Henzinger, Suñol, & Weber, 2010; Jacso, 2008a, 2008b, 2008c, 2008d; Levine-Clark & Gil, 2009; Meho & Rogers, 2008; Meho & Yang, 2007; Norris & Oppenheim, 2007;

* Correspondence address: Departamento de Metodología, Facultad de Psicología, Universidad Complutense, Campus de Somosaguas, 28223 Madrid, Spain. Tel.: +34 913 943 061; fax: +34 913 943 189.
E-mail address: miguel@psi.ucm.es

1751-1577/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.joi.2012.07.004

Vaughan & Shaw, 2008; Vieira & Gomes, 2009; Walters, 2009). The second problem is that these databases also return "phantom citations" (García-Pérez, 2011) defined by Jacso (2008a) as papers purported to cite a given paper when they actually do not. The only way around these difficulties is a painstaking process of multi-source retrieval and authentication (García-Pérez, 2010). Obviously, no modification of the *h* index and no alternative index will solve problems derived from incomplete coverage or inaccurate citation counts, but these problems might be regarded as random errors evenly distributed across individuals.

A further difficulty is encountered in the practical use of the *h* index and most of its scalar variants, namely, that many researchers turn up with the same *h* index even when their overall publication record and citation counts differ substantially. This feature lends these indices unusable when the goal is to rank individual applicants for funding or promotion. The situation is further complicated in non-mainstream areas where *h* indices tend to be low and where many researchers with qualitatively and quantitatively different careers have the same *h* index (García-Pérez, 2009a). Differences among researchers with the same *h* manifest in the number of excess citations received by core papers and also in the number and distribution of citations among non-core papers, which indicates that the top and the tail of the citation curve may be useful for distinguishing researchers with the same *h*. The tail of the citation curve includes relatively old papers that are unlikely to receive any further citations but it is also populated with recently published papers that may eventually receive sufficient citations to enter the core and, thus, the tail grows with the continued output of a researcher. On the other hand, the top of the citation curve grows when formerly tail papers enter the core, but also as core papers continue to receive citations even when the researcher has ceased to be productive. Thus, the tail carries information about recent contributions arising from a researcher's activity whereas the top carries information about the continued impact of the past activity of a researcher. Because the top of the citation curve thus indicates the continually growing impact of a researcher's output, it is remarkable that it has received so little attention thus far.

To fill this gap, this paper proposes and evaluates an extension of the *h* index based on a generalization of the multidimensional *h* index proposed by García-Pérez (2009b). In brief, the *h* index is also extended here to vector form by progressively computing new components not only toward the tail of the citation curve but also up its top. Other indices have recently been proposed that consider the top or the tail of the citation curve (Bornmann, Mutz, & Daniel, 2010; Dorta-González & Dorta-González, 2011; Egghe, 2010; Franceschini & Maisano, 2010; Kuan, Huang, & Chen, 2011; Ye & Rousseau, 2010) but the two-sided *h* index proposed here does it in a distinctly different way. The two-sided *h* index is defined in the next section, a subsequent section illustrates its capabilities to differentiate the accomplishments of researchers with the same *h* index and, finally, various criteria are discussed and exemplified that use the two-sided *h* index to rank researchers with the same or similar *h*.

2. The two-sided *h* index

Let N be the total number of papers published by a researcher and let $\boldsymbol{C}=(c_1, c_2, \ldots, c_N)$ be an N-dimensional vector of citation counts whose components indicate the number of citations received by each of those papers. The components of $\boldsymbol{C}$ are assumed ordered so that $c_i \geq c_{i+1}$ for all $1 \leq i < N$. With this notation, the *h* index is the largest i (for $1 \leq i \leq N$) satisfying $c_i \geq i$. Fig. 1 shows citation curves for two actual researchers, plotting the value of components of $\boldsymbol{C}$ against the component index. These curves illustrate a geometrical characteristic of the *h* index, namely, *h* is the length of the side of the largest square from the origin that can be fitted under the curve (gray-shaded area in Fig. 1). This area is sometimes referred to as the Durfee square (Anderson, Hankin, & Killworth, 2008; Prathap, 2010). The Durfee square is seen in Fig. 1 to leave much of the area under the citation curve uncovered, and also to be insensitive to potentially large differences in these uncovered areas across researchers with the same *h*. Papers comprising the *h* core typically received excess citations that are not represented in the *h* index and that define the shape of the top of the citation curve (i.e., the pink-shaded area above the Durfee square in Fig. 1). Similarly, the $N-h$ non-core papers have also received citations that are not represented in the *h* index either and that define the shape of the tail of the citation curve (i.e., the blue-shaded area on the right of the Durfee square in Fig. 1). Differences between the relative sizes of the top and the tail of the citation curve across researchers with the same *h* can be very large, as shown in Fig. 1.

The two-sided *h* index is illustrated by the additional squares represented under the citation curves in Fig. 1, which provide a tessellation by progressively filling the top and the tail of the citation curve with additional squares (indices) analogously defined. Formally, the two-sided *h* index of length k is defined as $h \pm k = (h_{-k}, \ldots, h_{-1}, h_0, h_1, \ldots, h_k)$, where negative subscripts refer to consecutive squares up the top of the citation curve whereas positive subscripts refer to consecutive squares out the tail of the citation curve. Indices with negative subscripts (i.e., h_{-j} for $j \geq 1$) are computed by finding the largest i (for $i \leq h_{-(j-1)}$) satisfying $c_i \geq i + \sum_{l=0}^{j-1} h_{-l}$ and, then, $h_{-j} = i$. Analogously, indices with positive subscripts (i.e., h_j for $j \geq 1$) are computed by finding the largest i (for $\sum_{l=0}^{j-1} h_l + 1 \leq i \leq N$) satisfying $c_i \geq i - \sum_{l=0}^{j-1} h_l$ and, then, $h_j = i - \sum_{l=0}^{j-1} h_l$. For the example in Fig. 1a, where $N = 28$ and $\boldsymbol{C} = (386, 282, 172, 113, 87, 83, 80, 69, 40, 38, 30, 28, 27, 24, 17, 14, 11, 11, 10, 7, 7, 4, 2, 1, 1, 1, 0, 0)$, the $h \pm 4$ index is (8, 8, 10, 12, **15**, 6, 2, 1, 1), where the scalar *h* index, denoted h_0 in the two-sided index, is printed in boldface to facilitate identification.

For researchers with the same scalar *h*, the two-sided *h* index thus defined renders patterns that vary within the two extremes discussed next for $h \pm 4$. In one extreme, consider a researcher who has published only *h* papers each of which has received more than 5*h* citations, which makes $h \pm 4 = (h, h, h, h, \boldsymbol{h}, 0, 0, 0, 0)$; in the other extreme, consider a researcher