Journal of Informetrics 5 (2011) 248–264

Contents lists available at ScienceDirect

Journal of Informetrics

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

A bibliometric index based on the collaboration distance between cited and citing authors☆

Maria Bras-Amorós[a,*], Josep Domingo-Ferrer[a], Vicenç Torra[b]

[a] Universitat Rovira i Virgili, Department of Computer Engineering and Maths, UNESCO Chair in Data Privacy, Av. Paso Catalans 26, E-43007 Tarragona, Catalonia, Spain
[b] IIIA-CSIC, Campus UAB, E-08193 Bellaterra, Catalonia, Spain

ARTICLE INFO

Article history:
Received 16 June 2010
Received in revised form 8 October 2010
Accepted 3 November 2010

Keywords:
Bibliometric indices
h-Index
c-Index

ABSTRACT

The popular *h*-index used to measure scientific output can be described in terms of a pool of evaluated objects (the papers), a quality function on the evaluated objects (the number of citations received by each paper) and a sentencing line crossing the origin, whose intersection with the graph of the quality function yields the index value (in the *h*-index this is a line with slope 1). Based on this abstraction, we present a new index, the *c*-index, in which the evaluated objects are the citations received by an author, a group of authors, a journal, etc., the quality function of a citation is the collaboration distance between the authors of the cited and the citing papers, and the sentencing line can take slopes between 0 and ∞. As a result, the new index counts only those citations which are significant enough, where significance is proportional to collaboration distance. Several advantages of the new *c*-index with respect to previous proposals are discussed.

1. Introduction

In the field of bibliometrics, the *h*-index by Hirsch (2005) has earned a lot of popularity, being publicized by Ball (2005) in *Nature* and implemented in the Web of Knowledge bibliometric database (Thomson's Scientific). Previous indicators were the total number of papers or the total number of citations. It is widely accepted that not all papers should count and Hirsch with his index suggested to count only those that were considered significant according to the number of citations.

However, less attention has been paid to citations. Just like not all papers count, neither all citations should count nor those that count should count in the same way. It is true that some attention has been devoted to differentiating self-citations, on the theoretical side by, *e.g.* Schreiber (2007), Derby (2008), Zhivotovsky and Krutovsky (2008), and, on the practical side, most notably by the CiteSeer database (CiteSeer.IST).

We suggest a new index following Hirsch's idea, but counting only those citations that are considered significant, where the significance of a citation is proportional to the collaboration distance between the cited and the citing authors.

The rest of this paper is organized as follows. Section 2 revisits the *h*-index and presents an abstraction of it leading to the *c*-index. Section 3 proposes the new index based on the collaboration distances between citing and cited authors. A

☆ This work was partly supported by the Spanish Government through projects TSI2007-65406-C03-01/02 "E-AEGIS", TIN2009-11689 "RIPUP", and CONSOLIDER INGENIO 2010 CSD2007-00004 "ARES", and by the Government of Catalonia under grants 2009 SGR 1135 and 2009 SGR 7. The second author is partly supported by the Government of Catalonia as an ICREA Acadèmia Researcher. The authors are with the UNESCO Chair in Data Privacy, but their views do not necessarily reflect those of UNESCO nor commit that organization.

* Corresponding author.
E-mail addresses: maria.bras@urv.cat (M. Bras-Amorós), josep.domingo@urv.cat (J. Domingo-Ferrer), vtorra@iiia.csic.es (V. Torra).

1751-1577/$ – see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.joi.2010.11.001

discussion of the advantages of the new index w.r.t. previous bibliometric indices is given in Section 4. Section 5 describes a procedure to efficiently compute the new index in practice. Experimental results are reported in Section 6. Section 7 contains some concluding remarks. Appendix A contains expanded experimental results.

2. From the *h*-index to the *c*-index

The h-index can be reinterpreted as follows. Consider the pool of all papers by an author. Consider the quality $Q(p)$ of a paper p to be the number of its citations. Sort all elements in the pool by decreasing order of their quality. That is, write the elements in the pool as $p_1, \ldots, p_n$, with $Q(p_i) \geq Q(p_{i+1})$. Then, draw the (decreasing) graph with all points $(i, Q(p_i))$. Let the *sentencing line* be the line through the origin with slope 1. The h-index is approximately the ordinate of the intersection of the sentencing line with the graph (see Fig. 1(a)). Formally, h is the maximum of the values $\min(i, Q(p_i))$ for $i \in \{1, \ldots, n\}$ (see Fig. 1(b)). That is,

$$h = \max\{\min(i, Q(p_i)) : i \in \{1, \ldots, n\}\} \tag{1}$$

In this way, there are h papers $p_1, \ldots, p_h$ with quality (citations) at least h, and the rest of papers with quality at most h.

From a formal point of view, Eq. (1) corresponds (see Torra & Narukawa, 2008 for details) to the Sugeno integral (Sugeno, 1974) of function Q with respect to a particular fuzzy measure. Namely, the measure, i in Eq. (1), roughly corresponds to the number of elements in the pool such that $Q(p) \geq Q(p_i)$.

Looking at the h-index in this more abstract way, we can see that it is defined by means of these three items:

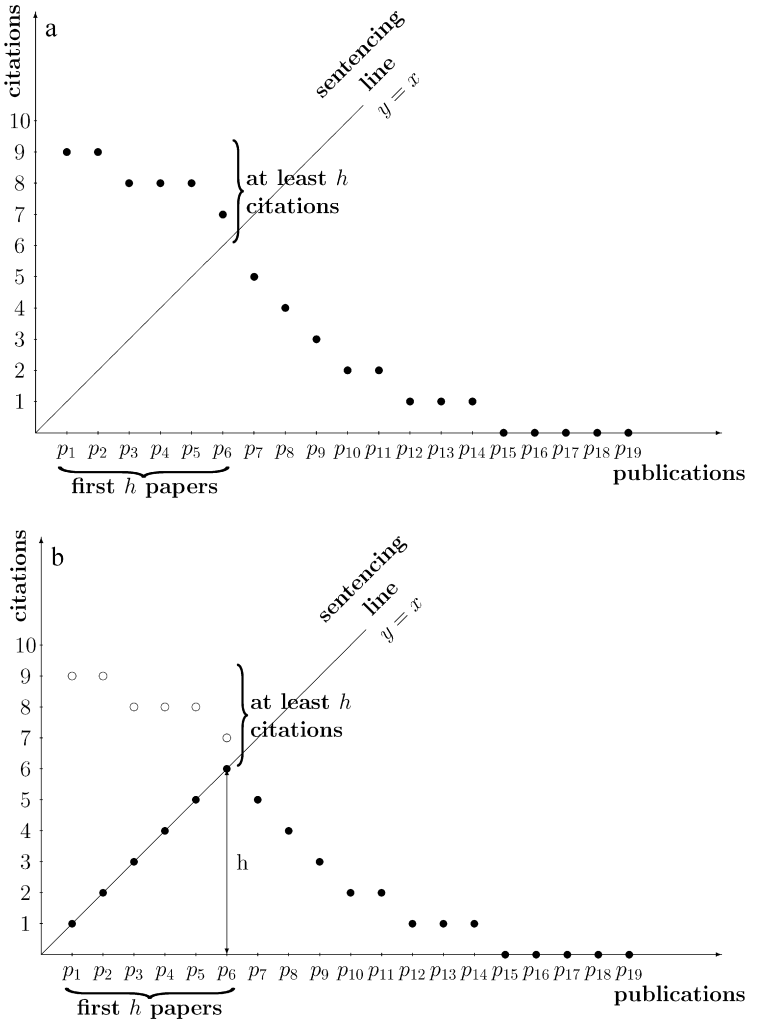

Fig. 1. (a) Reinterpretation of the h-index (inspired in Wikipedia). (b) Representation of $\min(i, Q(p_i))$.