Journal of Informetrics 7 (2013) 792–802

Contents lists available at ScienceDirect

Journal of Informetrics

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

Scientific impact assessment cannot be fair

Marek Gagolewski[a,b,*]

[a] Systems Research Institute, Polish Academy of Sciences, ul. Newelska 6, 01-447 Warsaw, Poland
[b] Faculty of Mathematics and Information Science, Warsaw University of Technology, ul. Koszykowa 75, 00-662 Warsaw, Poland

ARTICLE INFO

Article history:
Received 25 March 2013
Received in revised form 1 July 2013
Accepted 2 July 2013

Keywords:
Impact functions
Aggregation
Decision making
Preference modeling
Hirsch's h-index
Scientometrics
Bibliometrics

ABSTRACT

In this paper we deal with the problem of aggregating numeric sequences of arbitrary length that represent e.g. citation records of scientists. Impact functions are the aggregation operators that express as a single number not only the quality of individual publications, but also their author's productivity.

We examine some fundamental properties of these aggregation tools. It turns out that each impact function which always gives indisputable valuations must necessarily be trivial. Moreover, it is shown that for any set of citation records in which none is dominated by the other, we may construct an impact function that gives any a priori-established authors' ordering. Theoretically then, there is considerable room for manipulation in the hands of decision makers.

We also discuss the differences between the impact function-based and the multicriteria decision making-based approach to scientific quality management, and study how the introduction of new properties of impact functions affects the assessment process. We argue that simple mathematical tools like the h- or g-index (as well as other bibliometric impact indices) may not necessarily be a good choice when it comes to assess scientific achievements.

1. Introduction

Policy managers, decision makers, and scientists across all disciplines show great interest in the development of sensible, just, and transparent assessment methods of individual scientific achievements. At first glance, it may seem that the adoption of any mathematical formula puts an end to discretionary rankings. However, it was the introduction of one particular tool by J.E. Hirsch (2005) that brought new hopes for the fairness of the quality evaluation process. One of the most attractive features – enthusiastically received by the bibliometric community – of the h-index (and related indices) is that it expresses as a single number both the quality of individual papers, as well as the overall author's productivity.

However, some of the studies revealed that particular classes of scientific impact indices may easily be manipulated. One can think of at least two kinds of such devious influencing:

1. The first one occurs when a scientist tries to artificially improve his/her position in a ranking. For example, Bartneck and Kokkelmans (2011) as well as Zhivotovsky and Krutowsky (2008) note that the h-index may be "inflated" by a clever self-citation pattern.

* Correspondence address: Systems Research Institute, Polish Academy of Sciences, ul. Newelska 6, 01-447 Warsaw, Poland. Tel.: +48 22 38 10 393; fax: +48 22 38 10 105
E-mail address: gagolews@ibspan.waw.pl

1751-1577/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.joi.2013.07.001

2. The second kind occurs when a person in charge of choosing the assessment procedure decides to favor a predefined clique. For example, it has been indicated recently that the generalized Hirsch index is highly sensitive to the application of simple input data transformations, which may be prone to fraudulence (see Cena & Gagolewski, 2013b; Gagolewski & Mesiar, 2012).

The first class of manipulation techniques is related to input data being aggregated (Should we use citations from *Web of Knowledge* rather that *SciVerse Scopus*? Should self-citations be removed? Should we normalize a paper's citations with respect to the number its authors? Should journal quality measures be used to assess the quality of a paper? etc.). The second one, only loosely related to the former, is of our interest in this paper. It concerns the ranking/aggregation tool itself, and assumes that we are working on representative and valid data. In this perspective, the "*Google Scholar h*-index" is the same mathematical tool as the *h*-index normalized for the number of coauthors.

Literally hundreds of studies were performed to examine the behavior of the *h*-, *g*-, and similar indices on real-world data (see e.g. Alonso, Cabrerizo, Herrera-Viedma, & Herrera, 2009; Bornmann & Daniel, 2009; Egghe, 2010). Few of them also considered the analysis of indices' theoretical properties, for example from the axiomatic (e.g. Woeginger, 2008b) or probabilistic/statistical perspective (cf. Nair & Turlach, 2012 or Gagolewski, 2013b). It turns out that equivalent mathematical objects were already known in other scientific domains. For example, Torra and Narukawa (2008) showed that the *h*-index is a Sugeno integral with respect to a counting measure from the fuzzy/monotone measure theory. The others studied similar indices in the context of aggregation theory (cf. e.g. Grabisch, Marichal, Mesiar, & Pap, 2009).

However, the most fundamental questions concerning the aggregation methods of scientific output quality measures still remain open. Is the very nature of the assessment process such that it inevitably produces debatable results? If so, how to show it in a formal manner? When can we rely on the automatically generated valuations? On the other hand, in which cases is there room for manipulation and favoritism in the hands of decision makers?

The answer to these questions is crucial, because automated decision making is becoming more and more popular nowadays. It is still hoped that this form of assessment process may become the cure for not-rare cases of disappointment with the subjectivity of the "human factor".

The paper is organized as follows. Section 2 recalls the notion of an impact function and shows its connection to a particular (binary) preordering relation defined on the set of vectors representing citation records.

In Section 3 we study whether there exists a nontrivial impact function that gives us "noncontroversial" results in "disputable" cases. Moreover, we formally show how the introduction of additional properties modifies results of pairwise comparisons of citation records. It turns out that the most "sensitive" part of an impact function creation is the transformation of a preordering relation (in which there is still some room for indefiniteness) to a total preorder.

In Section 4 we explore the possibility of creating an impact function that generates an arbitrary, preselected ranking of a set of authors. Additionally, we present an illustration concerning a particular class of impact functions (a generalized *h*-index) applied on exemplary, real-world scientometric data set.

Finally, in Section 5 we discuss the implications of the results.

2. Impact functions

In order to study any real-world phenomenon, we have to establish its abstract *model*, preferably in the language of mathematics. Let us assume that some a priori chosen, reliable paper quality measure takes values in $\mathbb{I} = [0, \infty)$. These may of course be non-integers, for example when we consider citations of papers that are normalized with respect to the number of coauthors. Importantly, the values are not bounded from above (and thus cannot be sensibly transformed to a finite-length interval, e.g. [0, 1]).

Moreover, let $\mathbb{I}^{1,2,\cdots}$ denote the set of all sequences (of arbitrary length) with elements in $\mathbb{I}$, i.e. $\mathbb{I}^{1,2,\cdots} = \bigcup_{n=1}^{\infty}\mathbb{I}^n$. Thus, the whole information on an author's output is represented by exactly one element in $\mathbb{I}^{1,2,\cdots}$.

We are interested in constructing an aggregation operator, i.e. a function that maps $\mathbf{x} \in \mathbb{I}^{1,2,\cdots}$ to some $\mathsf{F}(\mathbf{x}) \in \mathbb{I}$, and which reflects the two following "dimensions" of an author's output quality:

(a) quality of his/her papers (his/her ability to write eagerly cited or highly valuated papers),
(b) his/her overall productivity.

Additionally, (c) we have no reason to treat any paper in a special way: only their quality measures should have impact on the results of evaluation, and not how they are ordered (e.g. by the number of coauthors, by time of publication, or how does the author want it). Of course, this is one of the many possible approaches, see e.g. (Grabisch et al., 2009) for the axiomatization of functions which only consider (a).

It is often assumed (see e.g. Franceschini & Maisano, 2011; Gagolewski & Grzegorzewski, 2011; Quesada, 2009, 2010; Rousseau, 2008; Woeginger, 2008a, 2008b, 2008c) that each *impact function* – i.e. an aggregation operator $\mathsf{F} : \mathbb{I}^{1,2,\cdots} \to \mathbb{I}$ to be applied in the assessment process that follows the above-mentioned characteristics – should *at least* be: