Journal of Informetrics 5 (2011) 565–573

Contents lists available at ScienceDirect

Journal of Informetrics

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

Impact functions on the citation network of scientific articles

José A. de la Peña[a,b,*]

[a] Instituto de Matemáticas, Universidad Nacional Autónoma de México, Mexico
[b] Centro de Investigación en Matemáticas, CIMAT, Guanajuato, Mexico

ARTICLE INFO

Article history:
Received 2 March 2010
Received in revised form 20 March 2011
Accepted 2 May 2011

Keywords:
Network of citations
Impact of papers
Eigenvalues of matrices associated to networks
Perron eigenvalue

ABSTRACT

Scientometric models, which consider papers in a undifferentiated way, are blind to important features of the citation network. We propose an approach for the definition of a function P_S, for any set of scientific articles S, which reflects global properties of the citation network associated to S. Such a function, that we propose as a measure of the impact of scientific papers, is constructed as solution of an iterated system of Perron-eigenvalue problems. We discuss differences with previously defined measures, in particular of the PageRank type.

1. Introduction

In the last years there has been increasing attention to citation-based statistics as a method to assess the quality of research of individuals, institutions and even countries. Many studies on the assessment of scientific production have appeared since the introduction of the impact factor of scientific journals in the 1960s, see for example (Adler, Ewing, & Taylor, 2008; Garfield, 1995, 1998). For individual scientists there have been attempts to represent the citation record by a single number, the most popular of such attempts being the *h*-index (Hirsch, 2006). For papers, the most common methods of assessment are either the simple count of citations or the sum of the citations weighted according to the impact factor of the journal of publication. In our view, these quantitative assessments ignore essential information contained in the network of citations.

Whatever it measures, we expect that a function ranking the scientific papers reflects rather an *intuitive* idea of quality than the quantity of the citations. Quantity should count only to differenciate between two papers cited by articles whose average "quality" is similar. In more dramatic terms, the ranking functions have to be concieved to cope with the paradigmatic question: which paper has more impact, one quoted by 10 unknown scientists or another just quoted by Einstein? On the other hand, an insight into the difference which entails the consideration of the "place" that an article occupies in the global network of citations vis–vis simply counting the number of citations may be gained by comparing the efficiency of the *G* oogle browser with the pre-Google search engines which counted the number of hits on a page as one of the main inputs for their ranking of web pages. As it is well known, Google search engine uses a Perron-type algorithtm (or Markov chain algorithm) for the ranking of a page by its role in the network of page citations. This is the direction that we explore in this paper for the consideration of the assessment of the impact of scientific articles. Moreover, when building a ranking function we shall recall that a not-so-unfrequent phenomenon is that a paper by a unknown scientist is discovered and

* Correspondence address. Instituto de Matemáticas, Universidad Nacional Autónoma de México, Mexico.
E-mail addresses: jap@matem.unam.mx, jap@cimat.mx

1751-1577/$ – see front matter
doi:10.1016/j.joi.2011.05.001

cited by a well-known scientist, the next generations of citations going rather to the known scientist than to the original source.

Taking the above remarks as a guide, we propose a number of properties ($P1$) to ($P6$) that a function has to fulfill in order to be considered as an *impact function* for scientific articles. We explicitely construct functions satisfying these properties by calculating, using Perron-Frobenius theory, the *weight* of the nodes (=articles) of a matrix associated to the citation network with some especific assigment of weights in the links. The motivation for the definition of impact functions as those satisfying ($P1$) to ($P6$) comes from network theory. Indeed, we shall expect an article to have a higher impact as its "position" in the network of citations becomes more prominent.

We introduce properties ($P1$) to ($P6$) in Section 2 and state the main mathematical results of this work. Section 3 is devoted to a general discussion on the assesment of scientific papers and the approach we propose. Among other issues, we discuss similarities to the Pinski and Narin approach (Pinski & Narin, 1976) and differences with previously defined measures, in particular of the PageRank type. Section 4 is devoted to explicit ad hoc examples. We present the proofs of theorems in Section 5.

2. Impact functions on the citation network

In our view, an *index of impact of scientific articles*, or simply, an *impact function*, for a given set of scientific papers S should be a non-negative valued function f satisfying the following properties:

($P1$): only articles a without citations have $f(a)=0$; for an article a cited by $b_1, \ldots, b_s$ we consider the average $\bar{f}(a):=AM\{f(b_1), \ldots, f(b_s)\} = 1/s\sum_{i=1}^{s} f(b_i)$, then we have:

($P2$): for any two articles a and a' cited, respectively, by $b_1, \ldots, b_s$ and $b'_1, \ldots, b'_t$ such that $\bar{f}(a) \leq \bar{f}(a')$ and $s \leq t$, then $f(a) \leq f(a')$;

($P3$): for an article a cited by $b_1, \ldots, b_s$ such that the average $\bar{f}(a) = 1$, then $f(a) \geq 1$ (resp. $f(a) \leq 1$) depending on whether the number of citations $c(a)=s$ is bigger or smaller than the average number of citations c_S of papers in S.

For the final properties, we consider S with the *network structure* induced by the citations, that is, we draw an arrow $a \to b$ whenever the paper $a \in S$ quotes the paper $b \in S$. Observe that, in this way, S becomes an oriented network without oriented cycles. For any $a \in S$ we denote by S_a the network induced from S by those papers which are proper predecessors of a in S.

($P4$): for any article $a \in S$, the network S_a determines the value $f(a)$, that is, for two articles a and a' if S_a and $S_{a'}$ are isomorphic networks, then $f(a)=f(a')$. Moreover,

($P5$): $f(a) = \sum_{b \to a} w(b)$ for some non-negative numbers $0 \leq w(b) = r_1 f(b) + r_2$ with $0 < r_1, r_2 < 1$ and every predecessor $b \to a$.

A subset S' of S with the induced network structure is said to be *terminal* in S if for every arrow $a \to b$ with $a \in S'$ then $b \in S'$. For a terminal subset S' of S and $a \in S'$ then S'_a is a terminal subset of S_a. For two articles $a, a' \in S$ we say that a' is *coterminal* to a, and write $a' \nabla a$, in case there is an injective function $h : S_{a'} \to S_a$ preserving the network structure and such that $h(a')=a$ and $h(S_{a'})$ is terminal in S_a. The last property we expect to be satisfied by an impact function refers to terminal substructures of S.

($P6$): suppose a' is coterminal to a in S, then $f(a') \leq f(a)$, with equality only in case a is coterminal to a'. Moreover, if $f(a) = \sum_{b \to a} w(b)$ (resp. $f(a') = \sum_{b' \to a'} w'(b')$) for some non-negative numbers $0 \leq w(b) = r_1 f(b) + r_2$ with $0 < r_1, r_2 < 1$ and every predecessor $b \to a$ (resp. $0 \leq w'(b') = r'_1 f(b') + r'_2$ with $0 < r'_1, r'_2 < 1$ and every predecessor $b' \to a'$) then $r_1 \leq r'_1$ and $r_2 \leq r'_2$.

According to ($P1$) and ($P4$), an impact function f is defined by induction on the order of paths in the oriented network S. Indeed, in the network of citations of S, papers just appearing can not be evaluated yet, hence their *impact* should be void, $f(a)=0$. If a paper a in S is cited by $b_1, \ldots, b_s$, the papers b_i have a later publication date than a, so we may assume that $f(b_i)$ are defined. Whatever the meaning of the impact values $f(b_i)$, the sum $\sum_{i=1}^{s} f(b_i) = sAM\{f(b_i) : 1 \leq i \leq s\}$ plays the main role for defining $f(a)$, as expressed by ($P2$) and ($P3$).

To fulfill property ($P3$) an estimate of c_S as a variable of time should be calculated for different disciplines of science. There is strong evidence that c_S differs from field to field of science but the growth of $c_S(t)$, for different disciplines along the time, shows some regularities, see for example (Adler et al., 2008; Yablonsky, 1980; www.eigenfactor.org/methods.htm) for a discussion of this topic. For instance, according to Garfield (1976), in 1976, c_S (1976) for the total of SCI publications, was between 12 and 13 and, no doubt, c_S (2011) is much higher.

As said in Section 1, the motivation to consider the properties of impact functions comes from network theory. Indeed, we shall expect an article a to have a higher impact as its "position" in the network of citations becomes more prominent, the value $f(a)$ depending only on the network S_a, as expressed by ($P4$) and ($P5$). Moreover, the fact that citations often do not refer to the original source, as for example when a rather unknown scientist is discovered and cited by a well-known scientist, the next generations of citations going rather to the known scientist (recall the cases of Ramanujan and Hardy in mathematics or Bose and Einstein in physics), calls for the consideration of higher-generation citations. But obviously, a second generation citation (not to say, third or higher) cannot "count" as a first generation citation (the real thing). Hence the consideration of *decay values* of the citations according to the generation is necessary. This is part of the content of property ($P5$): each predecessor $b \in S_a$ contributes to the value of $f(a)$ by $r_1 f(b) + r_2$ for numbers $0 < r_1, r_2 < 1$ independent of b.

In case an article a' is coterminal to another article a, we get a richer citation history of a and therefore $f(a') \leq f(a)$ is expected. On the other hand, if $h : S_{a'} \to S_a$ is a network preserving injection with $h(a')=a$, there are more generations of