An informetric model for the Hirsch-index

Leo Egghe
Hasselt University, Agoralaan, 3590 Diepenbeek, Belgium & Antwerp University, IBW, Universiteitsplein 1, 2610 Wilrijk, Belgium

Ronald Rousseau
KHBO (Association K.U.Leuven), Industrial Sciences and Technology, Zeedijk 101, 8400 Oostende, Belgium, & Hasselt University, Agoralaan, 3590 Diepenbeek, Belgium & Antwerp University, IBW, Universiteitsplein 1, 2610 Wilrijk, Belgium

Abstract

The h-index (or Hirsch-index) was defined by Hirsch in 2005 as the number h such that, for a general group of papers, h papers received at least h citations while the other papers received no more than h citations. This definition is extended here to the general framework of Information Production Processes (IPPs), using a source-item terminology. It is further shown that in each practical situation an IPP always has a unique h-index. In Lotkaian systems h = $T^{1/\alpha}$, where T is the total number of sources and α is the Lotka exponent. The relation between h and the total number of items is highlighted.

Key words and phrases: Hirsch-index, h-index, power laws, Information Production Process (IPP), sources and items

Introduction

The h-index has recently been defined by Jorge E. Hirsch (2005). A scientist, author or co-author of T articles has an h-index equal to h if h of his/her articles received at least h citations each, and the other T-h articles received each no more than h citations. The h-index is a single measure of visibility of a whole group of articles, incorporating publications as well as citations. It has an advantage over some other simple indicators such as 'number of significant papers' (which is arbitrary), or 'number of citations received by each of the q most-cited papers' (which is not a single number) (Hirsch, 2005).

The h-index has generally been well-received by the research community (Ball, 2005; Popov, 2005), and even in informetrics (Bornmann & Daniel, 2005; Braun, Glänzel & Schubert, 2005; Glänzel, 2006a). The authors of the latter article underline the fact that the h-index is robust in the sense that it is insensitive to an accidental excess of uncited articles, and to one or several extremely highly-cited articles. They also remark that the h-index combines the effects of quantity (number of publications) and quality (number of citations). Finally, they predict that Hirsch-type indexes will challenge scientometricians and informetricians. Of course, the h-index has also a number of disadvantages as pointed out by Van Raan (2005) and Glänzel (2006a). This makes this new indicator an addition, but certainly not a substitute for the more advanced indicators as used by institutes such as CWTS (Leiden, the Netherlands).

The authors certainly agree with the idea that the h-index deserves further study and hence will attract a lot of attention from the scientometric and informetric community. Accordingly they intend in this article to contribute to the mathematical study of the h-index, see also (Glänzel, 2006b).

In the next section we consider the h-index in the general context of source-item relations, hence placing it in a more general context than has been done so far. Then we prove an existence theorem for the h-index. Concretely, we show that any Information Production Process (in short: IPP) has a unique h-index, an important fact to know.

In a following section we calculate the h-index for so-called Lotkaian systems, systems governed by a power law, or stated otherwise, where Lotka's law is valid. We prove that if the system has T sources and a Lotka function with exponent α, then the h-index is equal to

$$h = T^{1/\alpha}$$

This relation is a concavely increasing power function in T, and a convexly decreasing function in α. A relation between h and the total number of items, denoted as A, is also determined.

Definition of the h-index and an existence theorem

Definition of the h-index

Consider an IPP consisting of sources and items (Egghe, 2005). This notion includes a plethora of source-item applications such as author-article (Lotka type) or journal-article (Bradford type) relations, including article-citation relations as considered by Hirsch (2005). Let g denote the rank-frequency function of such a system: if sources are ranked (using the symbol r) in decreasing order of production, then g(r) denotes the number of items 'produced' by the source at rank r. In the remainder of this article we will use a continuous approach since this is mathematically more convenient (Egghe, 2005). In this framework g(r) denotes an item density, where the function g is defined on [0,T] and assumed to be strictly positive, strictly decreasing and continuous.

Clearly, in this framework the h-index is defined as the value r such that

$$g(r) = r \qquad (1)$$

Indeed, since g is decreasing, equation (1), in its discrete interpretation, states that there are r sources with r or more items while the other have no more than r items. Also in the continuous case, equation (1) is taken as the definition of the h-index. Note that r can be described as a fixed point for the function g. We now show that each IPP has a unique h-index.

Existence theorem for the h-index

Theorem A. Each IPP as described above has a unique h-index.

Proof. Since g is a rank-frequency function it is strictly positive. We further assume that g(T) = 1 < T (the last source has production 1). Define now,

$$k(x) = g(x) - x \qquad (2)$$

The function k(x) is continuous and strictly decreasing on [0,T], k(0) > 0 and k(T) < 0. By the intermediate value theorem the function k takes all values on the interval [k(T) , k(0)] (where we have used the fact that k is decreasing). Hence there exists a unique value $x \in [0,T]$ such that k(x) = 0. Consequently, g(x) = x. Uniqueness follows from the fact that g is strictly monotonous (here decreasing). This value x is by definition the h-index.

In practical i.e., discrete, cases h is taken to be that rank r where g(r) ≥ r and g(r+1) < r+1. We finally note that also in the degenerate case T = 1 (just one source), g(1) = g(T). Hence, in the discrete case the h-index is equal to 1, so g(1) = 1.

The h-index in Lotkaian systems

Derivation of the h-index using power laws (Lotka's law)

In this framework we consider the size-frequency function f : $[1,\infty[\rightarrow]0,C]$ of the form

$$f(j)=\frac{C}{j^{\alpha}} \tag{3}$$

where C > 0 and α > 1 (Egghe, 2005). In a discrete setting f(j) denotes the number of sources with production j. In the continuous framework f is again interpreted as a density. We next prove a theorem characterizing the h-index for such power law systems. This theorem describes h as a function of the exponent α and the total number of sources T. This result has essentially also been obtained in (Glänzel, 2006b), by an argument using asymptotic behaviour, and discrete variables.

Theorem B

Suppose a Lotkaian system with T sources is given. Then the h-index is:

$$h=T^{1/\alpha} \tag{4}$$

Proof. In this framework the total number of sources with n or more items equals

$$\int_{n}^{\infty} f(j)\,dj=\int_{n}^{\infty}\frac{C}{j^{\alpha}}dj=\frac{C}{\alpha-1}\,n^{1-\alpha} \tag{5}$$

where α > 1. The total number of sources, T, is now equal to

$$T=\int_{1}^{\infty} f(j)dj=\frac{C}{\alpha-1} \tag{6}$$

Combining equations (5) and (6) yields that the number of sources with n or more items is equal to $T n^{1-\alpha}$. We conclude that the h-index, h, is equal to that number n such that $T n^{1-\alpha} = n$. Consequently

$$T = h^{\alpha} \tag{7}$$

or: $h = T^{1/\alpha}$. This proves Theorem B.

Corollary A

In the case of Lotka's square law, h = $\sqrt{T}$, a remarkable coincidence with Price's square law (Price, 1963). Note though that for technical reasons there is no real equivalence, see (Egghe, 2005, p. 217).

Derivation of the h-index using rank power laws (Zipf's law)

A result similar to that of Theorem B can be proved using Zipf's law, i.e. the rank-frequency function

$$g :]0, T] \to [1, +\infty[: r \to g(r) = \frac{B}{r^{\beta}} \tag{8}$$

with B, $\beta > 0$.

Theorem C. Suppose we have a system that can be described by equation (8). Then

$$h = B^{\frac{1}{\beta+1}} \tag{9}$$

Proof. The definition of the h-index: g(r) = r yields here: $g(r)=\frac{B}{r^{\beta}}=r \quad or \quad B=r^{\beta+1}$.

Hence:

$$h=r=B^{\frac{1}{\beta+1}}$$

This proves Theorem C.

Note that rank-frequencies can also be studied using the Mandelbrot formalism. The h-index is then the solution of $g(r)=\frac{M}{(1+Nr)^{\beta}}=r$. This equation can, in general, only be solved in a numerical way (the case β = 1, being an obvious exception). For this reason we do not consider it anymore.

Equivalence of the results obtained in Theorems B and C

Proposition A (Egghe, 2005; Exercise II.2.2.6)

The following assertions are equivalent

(i) $f(j)=\frac{C}{j^{\alpha}}$, with C > 0, α > 1 (constants) and $j \in [1,+\infty[$.

(ii) $g(r)=\frac{B}{r^{\beta}}$, with B, β > 0 (constants) and $r \in]0,T]$, where T denotes the total number of sources.

Moreover, the relations between the parameters are:

$$B=\left(\frac{C}{\alpha-1}\right)^{\frac{1}{\alpha-1}} \quad or \quad C=B^{(\alpha-1)}(\alpha-1) \tag{10}$$

$$\beta=\frac{1}{\alpha-1} \quad or \quad \alpha=\frac{1+\beta}{\beta} \tag{11}$$

For the reader's convenience a proof is provided in the appendix.

Equations (6), (10) and (11) clearly show that Theorems B and C are equivalent. In our opinion Theorem B and in particular equation (4) is the most appealing result since it relates the h-index directly to the total number of sources and Lotka's α.

In the next section we prove some further properties of the h-index.

Further results on the h-index

The following corollary gives more details on the relation between the h-index, the Lotka exponent and the total number of sources.

Corollary B

(i) The h-index is a concavely increasing function of the total number of sources T (keeping the Lotka exponent α constant).

(ii) The h-index is a convexly decreasing function of the Lotka exponent α (keeping the total number of sources constant, T > 1).

Proof. (i) is evident since α > 1, while (ii) follows easily from a straightforward calculation.

Note that in Corollary B (i) it is allowed to keep the Lotka exponent α constant while T is variable, and that in (ii) T can be kept constant while the Lotka exponent α is variable because there is a third parameter, namely C, involved (see equation (6)). For more information on the existence of certain Lotka laws the reader is referred to Egghe (2005, II.2.1).

We can also show a functional relation between the h-index h and the total number of items A. This is done in the following corollary.

Corollary C

In a Lotkaian system with A items and α > 2, the h-index is equal to

$$h=\left(\frac{\alpha-2}{\alpha-1}\,A\right)^{1/\alpha} \tag{12}$$

Proof. It is clear that $A=\int_1^\infty j\,f(j)\,dj=\int_1^\infty \frac{C}{j^{\alpha-1}}\,dj$, from which it follows that $A=\frac{C}{\alpha-2}$.

Hence, by (6):

$$A=T\,\frac{\alpha-1}{\alpha-2} \tag{13}$$

Substituting equation (13) in equation (4) yields the result. Note that $\frac{\alpha-1}{\alpha-2}=\frac{A}{T}$ = μ, the average number of items per source.

Remark. In this article we have considered the case $j\in[1,+\infty[$, implying an unrestricted item density. It is, however, also possible to study our results in a framework of a finite maximum item density, cf. (Egghe, 2005, Chapter II). This is left as an exercise.

Summary and conclusions

We have considered the Hirsch-index, h, in the context of Information Production Processes (IPPs) and shown an existence theorem in this general framework. In Lotkaian systems h = $T^{1/\alpha}$, where T is the total number of sources, and α is the Lotka exponent. The relation between these parameters is further highlighted. In (Egghe, 2006) we will continue our study of the h-index by studying the dynamics of the h-index in a time dependent framework. Clearly the theoretical study and practical applications of the h-index will occupy information scientists for many years to come.

References

BALL, P. (2005), Index aims for fair ranking of scientists, *Nature*, 436, p. 900.

BORNMANN, L., DANIEL, H.-D. (2005), Does the h-index for ranking of scientists really work? *Scientometrics*, 65, 391-392.

BRAUN, T., GLÄNZEL, W., SCHUBERT, A. (2005), A Hirsch-type index for journals, *The Scientist*, 19(22), p.8.

EGGHE, L. (2005), *Power Laws in the Information Production Process: Lotkaian Informetrics*. Elsevier, Oxford (UK).

EGGHE, L. (2006). Dynamic h-index: the Hirsch index in function of time. Preprint, Hasselt University.

GLÄNZEL, W. (2006a), On the opportunities and limitations of the H-index. *Science Focus*, 1(1), 10-11 (In Chinese).

GLÄNZEL, W. (2006b), On the H-index – a mathematical approach to a new measure of publication activity and citation impact. *Scientometrics,* 67(2). In press.

HIRSCH, J. E. (2005), An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, 102, 16569-16572.

POPOV, S.B. (2005), A parameter to quantify dynamics of a researcher's scientific activity. arXiv:physics/0508113

PRICE, D.J. de SOLLA (1963), *Little Science, Big Science*. Columbia University Press, New York.

VAN RAAN, A.F.J. (2005), Comparison of the Hirsch-index with standard bibliometric indicators and with peer judgment for 147 chemistry research groups, arXiv:Physics/0511206.

Appendix: proof of Proposition A

Proposition A

The following assertions are equivalent

(i) $f(j)=\frac{C}{j^{\alpha}}$, with C > 0, α > 1 (constants) and j $\in[1,+\infty[$.

(ii) $g(r)=\frac{B}{r^{\beta}}$, with B, β > 0 (constants) and r $\in]0,T]$, where T denotes the total number of sources.

Moreover, the relations between the parameters are:

$$B=\left(\frac{C}{\alpha-1}\right)^{\frac{1}{\alpha-1}} \quad or \quad C=B^{(\alpha-1)}(\alpha-1) \tag{10}$$

$$\beta=\frac{1}{\alpha-1} \quad or \quad \alpha=\frac{1+\beta}{\beta} \tag{11}$$

Proof. The basic relation between the functions f and g is given as:

$$r=g^{-1}(j)=\int_{j}^{\infty} f(s)\,ds \tag{14}$$

Indeed: for j = g(r), we have that r = g^{-1}(j). Here g^{-1} denotes the inverse function of g, its existence following from the fact that g is strictly decreasing, hence injective. The symbol r denotes the number of sources with item density larger

than or equal to j, and this is also the case for the right hand side of equation (14), cf. (Egghe, 2005, Chapter II). By equation (14) we further see that

$$f(j) = -\frac{1}{g'\left(g^{-1}(j)\right)} \tag{15}$$

Suppose that part (i) is given, then (14) yields: $r = g^{-1}(j) = \int_j^{\infty} \frac{C}{s^{\alpha}} ds = \frac{C}{\alpha - 1} j^{1-\alpha}$,

hence j = g(r) = $\left(r \frac{\alpha - 1}{C} \right)^{\frac{1}{1-\alpha}}$ or g(r) = $\frac{C^{\frac{1}{\alpha-1}}}{\left(r(\alpha-1)\right)^{\frac{1}{\alpha-1}}}$. Hence

$$g(r) = \frac{B}{r^{\beta}} \tag{16}$$

with B and β as given in equations (10) and (11).

Conversely, if (ii) is given then, then it follows that g'(r) = - β B $r^{-(\beta+1)}$, and hence, as j = g(r), it follows from equation (16) that $r = \left(\frac{B}{j}\right)^{1/\beta}$. Equation (15) then yields: $f(j) = -\frac{1}{-\beta B r^{-(\beta+1)}} = \frac{1}{\beta B \left(\frac{B}{j}\right)^{-\frac{\beta+1}{\beta}}} = \frac{\frac{B^{1/\beta}}{\beta}}{j^{\frac{\beta+1}{\beta}}}$.

This expression is of the form f(j) = $\frac{C}{j^{\alpha}}$, with $\alpha = \frac{\beta+1}{\beta}$ and $C = \frac{B^{1/\beta}}{\beta} = (\alpha - 1) B^{(\alpha-1)}$.

This proves the proposition.