The Hirsch-index of set partitions

by

L. Egghe

Universiteit Hasselt (UHasselt), Campus Diepenbeek, Agoralaan, B-3590 Diepenbeek, Belgium (*)
and
Universiteit Antwerpen (UA), Stadscampus, Venusstraat 35, B-2000 Antwerpen
leo.egghe@uhasselt.be

ABSTRACT

The Hirsch-index (h-index) is calculated on citations that papers (e.g. of authors or journals) receive. Hence we can consider the h-index as calculated on a partition of the same set of citations.

In this paper we will study the h-index, dependent on the particular partition of this set. We will do this in the discrete case as well as in a continuous Lotkaian setting.

In the discrete setting we will determine h-indices of successive refinements of partitions. We show that the corresponding h-indices do not form a monotonic sequence and we determine the maximal value of an h-index in such a system.

In the continuous Lotkaian setting we prove that, given a set of citations of cardinality A, the h-index only depends on the average number of citations that an author or a journal receives. This functional dependence is calculated and we show that it has a unique maximum for which formulae are given. This is the highest possible h-index, given a set of citations of fixed cardinality. Examples confirm the theory.

(*) Permanent address
Key words and phrases: partition, Hirsch- index, h-index, average.

Introduction

The description of the Hirsch-index (h-index) is well covered in the literature (see e.g. the review article Egghe (2010), covering the literature up to (and including) 2008). It can be applied to various situations; in the original article of Hirsch (Hirsch (2005)) it was applied to authors who's papers receive citations: then h is the largest rank $r = h$ (where papers are arranged in decreasing order of the number of received citations) such that all papers on ranks $1, 2, ..., h$ receive at least h citations. Then h is called the h-index of this author.

The h-index has attracted numerous applications (again see Egghe (2010)): h-index of journals, based on citations received by its papers (Braun, Glänzel and Schubert (2006)), h-index of topics (again based on citations received by papers on this topic) (Banks (2006)), groups of authors (institutes) (van Raan (2006), Egghe and Rao (2008)), h-indices based on downloads (instead of citations) (O'Leary (2008), Hua, Rousseau, Sun and Wan (2009)), and so on (see Egghe (2010) for many more examples and references).

In the case of citations to authors and journals, one could consider the set of citations as being partitioned over the cited authors as well as over the cited journals. This yields two different partitions of the same "citations set" and hence two different h-index values.

Another example is given in Kim, Lee and Park (2009), based on ideas in Lin and Rousseau (2009), is as follows. A collection of books (e.g. in a library) can be evaluated based on its borrowings: the h-index can be calculated on the (decreasing) list of the number of borrowings of the books. Alternatively, one can use the same borrowings as distributed over the set of borrowers of these books and again calculate the (different) h-index. The former h-index is called the "collection-side h-index" and the latter h-index is called the "user-side h-index) (Kim, Lee and Park (2009)).

One can generalise this situation as follows. We have a set of citations (more generally a set of items). This set is partitioned, yielding a set of subsets which serve as sources producing the corresponding items (see Fig. 1 for an example).

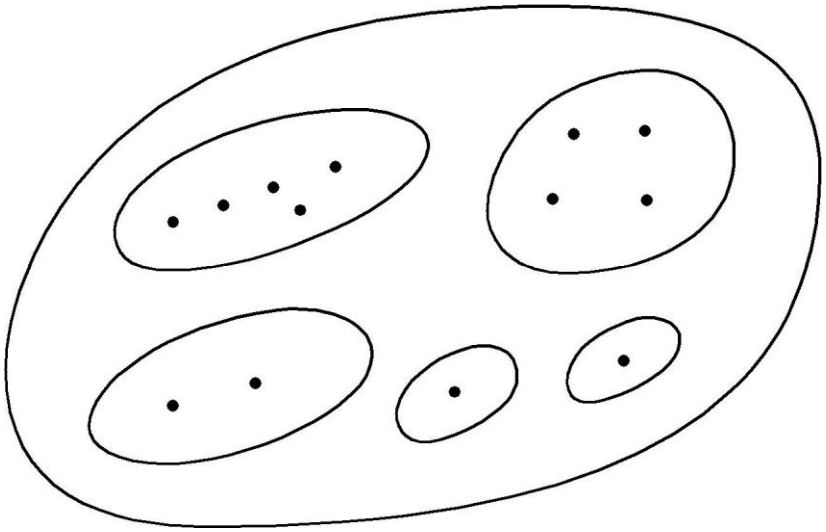

Fig. 1 An item set of 13 items, partitioned over 5 sources with number of items (in decreasing order) $5,4,2,1,1$ yielding an h-index $h=2$.

Putting the number of items in each subset in decreasing order, we can calculate the h-index of this situation ($h=2$ in the case of Fig. 1).

Of course, given an item set, one can partitionate this set in many ways. Each time we can calculate the h-index of this partition. The distribution of the h-indices is the topic of this paper.

The next section deals with the discrete case (as depicted in Fig. 1). More in particular we study the h-index sequence when we study partitions of the item-set from fine to rough as in the case of clustering item points: starting from all points forming separate clusters until we have clustered all points into one cluster (each cluster is the same as a source containing the corresponding items). We prove that the h-index sequences are (in general) non- monotonic and start and end in $h=1$. All sequences have length M (where M is the cardinality of the item-set) and the highest possible h-index is $h=\left[\sqrt{M}\right]$, where $\left[\sqrt{M}\right]$ denotes the largest entire number smaller than or equal to $\sqrt{M}$. Also a note is given on the number of possible linear order sequences, given an item-set with M elements.

In the continuous setting a partition of the fixed item-set is replaced by a Lotkaian system. We show that the h-index is only dependent on the average number μ of items per source. We will prove that the function $h(\mu)$ equals

$$h(\mu) = \left(\frac{A}{\mu}\right)^{\frac{\mu-1}{2\mu-1}} \tag{1}$$

where A denotes the total number of items (a fixed number). We prove that this function has a unique maximum for $\mu = \mu_0$ satisfying

$$\ln\left(\frac{A}{\mu_0}\right) = 2\mu_0 + \frac{1}{\mu_0} - 3 \tag{2}$$

and that this maximal h-index is given by (1) for $\mu = \mu_0$, which is equal to

$$h(\mu_0) = e^{\frac{(\mu_0-1)^2}{\mu_0}} \tag{3}$$

These results are confirmed in all examples presented. The paper closes with some conclusions and open problems.

Discrete case

Let us give an example in case our item-set contains 5 elements. We have a situation as depicted in Fig. 2.

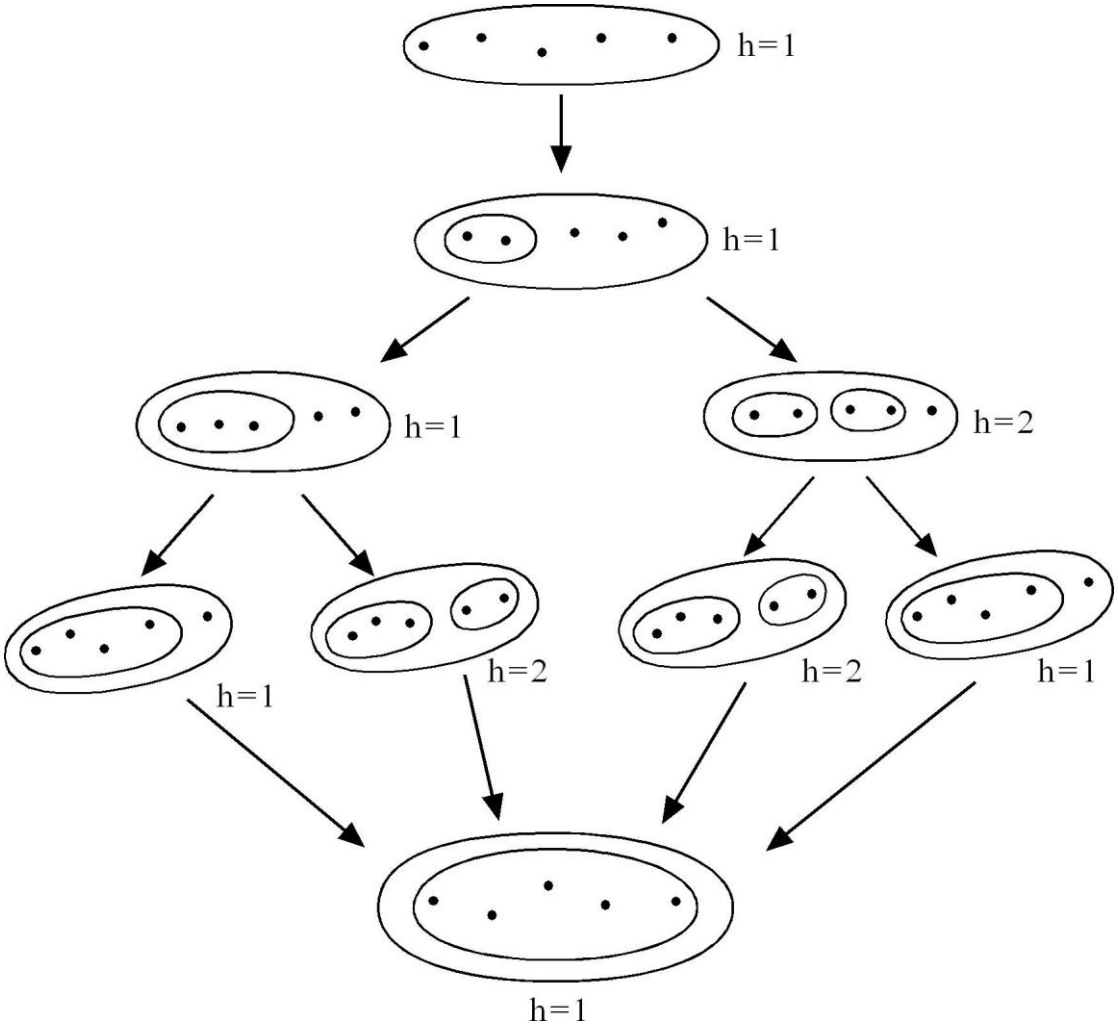

Fig. 2 Partitioning an item-set containing 5 elements and corresponding h-indices.

In the above example, all points are unlabelled. We did not label the points (as e.g. {1,2,3,4,5}) since this is irrelevant for the calculation of the h-index. The process shown in Fig. 2 follows the clustering method: going from the finest partition (5 clusters, each containing 1 point) to the roughest cluster (1 cluster containing 5 points).

In terms of concentration theory (see e.g. Egghe (2005), Chapter IV) we start from complete equality (lowest concentration): 5 sets each with 1 point to highest concentration: 1 set with all points (and the other sets are empty).

If the starting set has M points, then the length of each sequence is M. This is clear since there are M points and since we cluster one step at a time: the i^{th} step $(i = 1, 2, ..., M)$ (where

$i=1$ denotes the starting set) yields a set of $M-i+1$ clusters, hence we end at $M-i+1=1$ or $i=M$ steps.

The maximum value of the h-index is $h_{max}=\left[\sqrt{M}\right]$, where $\left[\sqrt{M}\right]$ denotes the largest entire number smaller than or equal to $\sqrt{M}$. Indeed, by the very definition of the h-index, the first h sources have at least h items so that $h^2 \le M$, hence $h \le \sqrt{M}$, and since h is an entire number, we have $h \le \left[\sqrt{M}\right]$.

Fig. 2 yields the following 4 h-index sequences: $(1,1,1,1,1)$, $(1,1,1,2,1)$, $(1,1,2,2,1)$ and $(1,1,2,1,1)$. So not all sequences are symmetric or monotone. Each sequence starts and ends in $h=1$ (even the second h-value is always $h=1$). We leave it to the reader to check the h-index sequences for $M=1,2,3,4$. For $M=6$ we have 11 (not necessarily different) h-index sequences, derived in the same was as in Fig. 2: $(1,1,1,1,1,1)$, $(1,1,1,1,2,1)$, $(1,1,1,2,1,1)$, $(1,1,1,2,2,1)$ (2x), $(1,1,2,2,2,1)$ (3x), $(1,1,2,2,1,1)$, $(1,1,2,1,1,1)$ and $(1,1,2,1,2,1)$.

The last sequence shows that the h-index can increase and decrease more than once. We leave open to calculate the number of (different) h-index sequences that are possible, given M. We have a solution for the labelled case (i.e. the elements in the item-set are numbered) but not in the unlabelled case as we deal with here. If all elements are labelled, then the first clustering step has $\binom{M}{2}$ possibilities, the second step has $\binom{M-1}{2}$ possibilities and so on. Altogether there are

$$\binom{M}{2}\binom{M-1}{2}\cdots\binom{4}{2}\binom{3}{2}\binom{2}{2} \tag{4}$$

possible sequences. A short calculation shows that (4) is equal to

$$\frac{M}{2^{M-1}}\prod_{i=1}^{M-2}(M-i)^2 \tag{5}$$

Continuous Lotkaian case

Now the item-set is an interval $[0, A]$, where A denotes the "total number" of items.

A partition (as in the previous section) is replaced by a Lotkaian system given by

$$f(j) = \frac{C}{j^{\alpha}} \tag{6}$$

$C > 0$, $\alpha > 1$, $j \geq 1$ (but in the sequel we will have to suppose $\alpha > 2$), where $f(j)$ is the density of sources with item density j. Hence

$$A = \int_{1}^{\infty} j f(j) dj = \frac{C}{\alpha - 2} \tag{7}$$

if $\alpha > 2$. Then (already if $\alpha > 1$)

$$T = \int_{1}^{\infty} f(j) dj = \frac{C}{\alpha - 1} \tag{8}$$

is the total number of sources (replacing the sets in the partition in the previous section).

It follows (cf. also Egghe (2005), Chapter II), that

$$\mu = \frac{A}{T} = \frac{\alpha - 1}{\alpha - 2} \tag{9}$$

is the average number of items per source. The different partitions (of the previous section) are now expressed by different Lotka exponents α.

If we have a Lotkaian system as above it was shown in Egghe and Rousseau (2006) that the h-index of such a system equals

$$h = T^{\frac{1}{\alpha}} \tag{10}$$

In this formula both α and T are variable (T= total number of sources, which were the sets in the partition of the item-set in the previous section). However (9) and (10) yield (the formula already appears in Egghe and Rousseau (2006))

$$h = \left(A \frac{\alpha - 2}{\alpha - 1} \right)^{\frac{1}{\alpha}} \tag{11}$$

$$h = \left(\frac{A}{\mu} \right)^{\frac{1}{\alpha}} \tag{12}$$

The latter expression can further be developed to become a pure function of μ:
by (9) we have

$$\alpha = \frac{2\mu - 1}{\mu - 1} \tag{13}$$

(note that, by (9), $\mu \in]1, +\infty[$). (13) in (12) yields

$$h = h(\mu) = \left(\frac{A}{\mu} \right)^{\frac{\mu - 1}{2\mu - 1}} \tag{14}$$

so $h(\mu)$ is a pure function of μ since the item-set is fixed and hence also A. We have the following main result of this paper.

Theorem 1

In case we fix the item-set such that A>1 and using the Lotkaian framework, then (14) expresses all possible values of the h-index. This function has a unique maximum in value $\mu = \mu_0$, satisfying $\varphi(\mu_0) = 0$, where

$$\varphi(\mu) = \ln\left(\frac{A}{\mu} \right) - 2\mu - \frac{1}{\mu} + 3 \tag{15}$$

The highest possible h-index is then given by (14) for $\mu = \mu_0$ but $h(\mu_0)$ also satisfies the simpler

$$h(\mu_0) = e^{\frac{(\mu_0 - 1)^2}{\mu_0}} \tag{16}$$

Furthermore $\lim_{\mu\to+\infty} h(\mu)=0$ and $\lim_{\mu\to 1_{>}} h(\mu)=1$.

Proof:

Formula (14) was proved above. It is clear that all values $\mu\in\left]1,+\infty\right[$ are possible since α can be any value >2 and by (9). We have

$$h'(\mu)=\left(\frac{A}{\mu}\right)^{\frac{\mu-1}{2\mu-1}}\frac{1}{(2\mu-1)^2}\left[\ln\left(\frac{A}{\mu}\right)-2\mu-\frac{1}{\mu}+3\right] \tag{17}$$

$$h'(\mu)=f(\mu)\varphi(\mu) \tag{18}$$

where

$$f(\mu)=\left(\frac{A}{\mu}\right)^{\frac{\mu-1}{2\mu-1}}\frac{1}{(2\mu-1)^2}>0 \tag{19}$$

and

$$\varphi(\mu)=\ln\left(\frac{A}{\mu}\right)-2\mu-\frac{1}{\mu}+3 \tag{20}$$

Now for $\mu\underset{>}{\to}1$ we have that $\varphi(\mu)\to\ln(A)>0$, since $A>1$ and for $\mu\to+\infty$ we have that $\varphi(\mu)\to-\infty$. Furthermore, $\varphi(\mu)$ is strictly decreasing (since $\ln\left(\frac{A}{\mu}\right)$ decreases strictly and since $2\mu+\frac{1}{\mu}$ increases strictly).

Finally, $\varphi(\mu)$ is also continuous. This implies that there is a unique value $\mu=\mu_0$ such that $\varphi(\mu_0)=0$ and hence $h'(\mu_0)=0$.

Now $\lim_{\mu\to+\infty} h(\mu)=0$ and $\lim_{\mu\to 1_{>}} h(\mu)=1$ as is readily seen. Since $A>1$ we can take $\mu=\frac{A+1}{2}>1$ for which $\frac{A}{\mu}>1$ and hence $h(\mu)>1$. Hence, since $h'(\mu_0)=0$ we have that $h(\mu)$ has a (unique) maximum in $\mu=\mu_0$.

So, the highest possible value of the h-index in such a system equals

$$h(\mu_0) = \left(\frac{A}{\mu_0} \right)^{\frac{\mu_0 - 1}{2\mu_0 - 1}} \tag{21}$$

But, since μ_0 satisfies $\varphi(\mu_0) = 0$ and by (15) we have

$$h(\mu_0) = \left(e^{2\mu_0 + \frac{1}{\mu_0} - 3} \right)^{\frac{\mu_0 - 1}{2\mu_0 - 1}}$$

$$h(\mu_0) = e^{\frac{(\mu_0 - 1)^2}{\mu_0}} > 1$$

which is readily seen. This ends the proof of this theorem.

Comments:

The above Theorem shows that $h(\mu)$ starts low $\left(\mu \underset{>}{\to} 1 \right)$ and ends low $(\mu \to +\infty)$ and that it has its maximal value in a certain "intermediate" value of $\mu = \mu_0$. Although the discrete case is different from this continuous Lotkaian situation, this result is in the same line with the h-index sequences obtained in the previous section. Hence, the theory developed here is an explanation for this phenomenon. We will now give some concrete examples which confirm these findings and we will also give the corresponding values of μ_0 and $h(\mu_0)$.

The following Table gives the values of μ_0, $h(\mu_0)$ calculated via (14) and $h(\mu_0)$ calculated via (16) (both values of $h(\mu_0)$ are very close: the small difference occurs since the value of μ_0 is rounded-off to 3 decimals), for various values of A. The calculations are done using the MATHCAD4.0 software.

Table 1: Values of μ_0, $h(\mu_0)$

A	μ_0	$h(\mu_0)$ (via (14))	$h(\mu_0)$ (via (16))
100	3.078	4.067	4.067
200	3.391	5.398	5.397

300	3.575	6.390	6.390
400	3.706	7.212	7.213
500	3.807	7.926	7.922
600	3.891	8.565	8.565
700	3.961	9.148	9.147
800	4.022	9.686	9.686
900	4.076	10.189	10.189
1,000	4.124	10.662	10.660
2,000	4.442	14.403	14.398
3,000	4.629	17.202	17.202
4,000	4.761	19.526	19.512
5,000	4.864	21.550	21.533
6,000	4.949	23.364	23.361
7,000	5.020	25.020	25.008
8,000	5.082	26.552	26.543
9,000	5.137	27.983	27.985
10,000	5.186	29.330	29.336

It is clear and logical that μ_0 and $h(\mu_0)$ increase (concavely) with A and that $h(\mu_0)$ increases with μ_0 for varying A (note that we proved in Theorem 1 that $h(\mu)$ is not an increasing function of μ but that it has a unique maximum in $\mu = \mu_0$). In Fig. 3 one can see some examples of the $h(\mu)$ functionality.

The function $h(\mu)$ for A=100; 500; 1,000 and 10,000 are depicted. They all have the same shape: start increasing convexly (although the convexity only becomes apparent for the higher values of A), reaches its unique top concavely and then, after a while, starts slowly decreasing convexly.

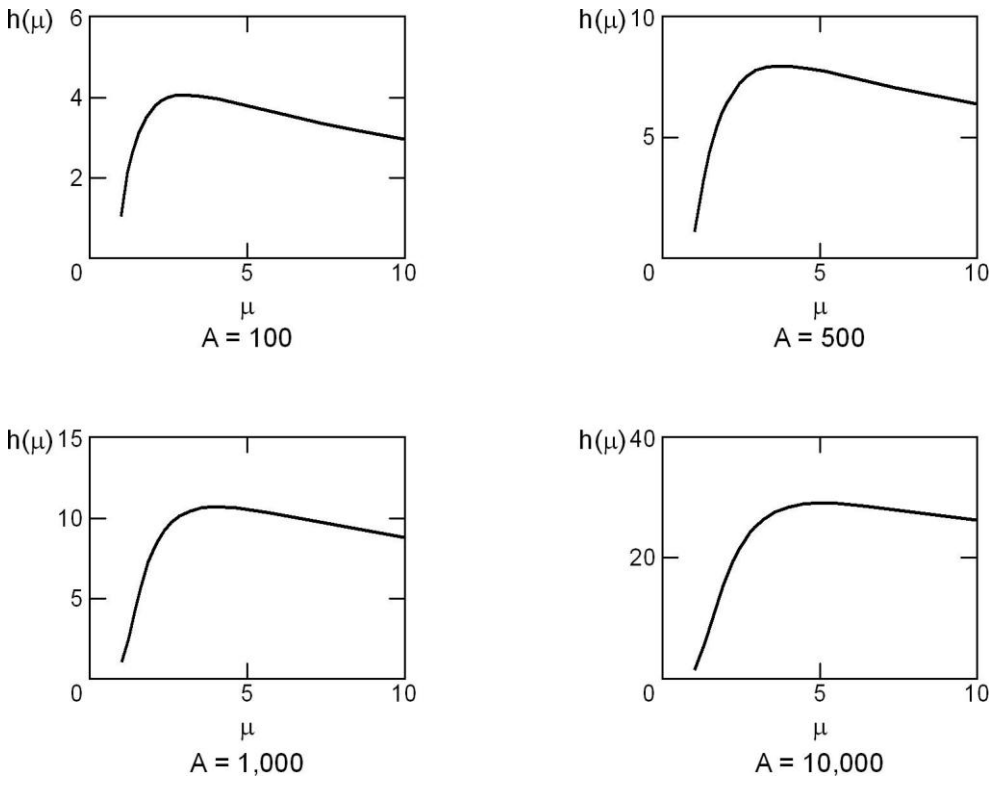

Fig.3 The function $h(\mu)$ in function of the average μ for various values of A.

Conclusions and open problems

This paper studies the different h-index values one can obtain when partitioning a fixed item-set. In the discrete case, the iterative procedure of clustering (one by one) from the finest (starting) case to the roughest case (all points in one cluster) yields h-index sequences of length M and with h-values at most $\left[\sqrt{M}\right]$. The highest values are somewhere in the interior of the sequence (each sequence starts and ends in $h=1$). We leave open how many unlabelled sequences can be formed.

In the continuous case, the item-set is an interval $[0, A]$ and partitions are replaced by a Lotkaian system where Lotka's α varies to yield different systems, comparable with the different partitions in the discrete case. We note that, since A is fixed, the h-index is only a function of μ, the average number of items per source. This function is studied: it increases until its unique maximum and then decreases. Its maximum is given by (16) in the point $\mu = \mu_0$ given by (15).

The conclusion of the discrete as well as of the continuous case is that the h-index is highest for an optimal spread of the items over the sources (which is also intuitively clear).

This paper offers a methodology to study other impact measures on set partitions (or its continuous variant). This will be done in a forthcoming paper.

References

M. G. Banks (2006). An extension of the Hirsch index: Indexing scientific topics and compounds. Scientometrics 69(1), 161-168.

T. Braun, W. Glänzel and A. Schubert (2006). A Hirsch-type index for journals. Scientometrics 69(1), 169-173.

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

L. Egghe (2010). The Hirsch index and related impact measures. Annual Review of Information Science and Technology, Volume 44 (B. Cronin, ed.), 65-114, Information Today, Inc., Medford, New Jersey, USA.

L. Egghe and I. K. R. Rao (2008). Study of different h-indices for groups of authors. Journal of the American Society for Information Science and Technology 59(8), 1276-1281.

L. Egghe and R. Rousseau (2006). An informetric model for the Hirsch-index. Scientometrics 69(1), 121-129.

J. E. Hirsch (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(46), 16569-16572.

J.-K. Wan, P.-H. Hua, R. Rousseau and X.-K. Sun (2010). The journal download immediacy index (DII): experiences using a Chinese full-text database. Scientometrics 82(3), 555-566.

P. J. Kim, J. Y. Lee and J.-H. Park (2009). Developing a new collection-evaluation method: mapping and the user-side h-index. Journal of the American Society for Information Science and Technology 60(11), 2366-2377.

Y. Liu and R. Rousseau (2009). Properties of Hirsch-type indices: The case of library classification categories. Scientometrics 79(2), 235-248.

D. E. O'Leary (2008). The relationship between citations and number of downloads in Decision Support Systems. Decision Support Systems 45(4), 972-980.