Item-time-dependent Lotkaian informetrics and applications to the calculation of the time-dependent h- and g-index
Peer-reviewed author version

Item-time-dependent Lotkaian informetrics

and applications to the calculation of the

time-dependent h- and g-index

by

L. Egghe

Universiteit Hasselt (UHasselt), Campus Diepenbeek, Agoralaan, B-3590 Diepenbeek, Belgium[1]

and

Universiteit Antwerpen (UA), Campus Drie Eiken, Universiteitsplein 1, B-2610 Wilrijk, Belgium

leo.egghe@uhasselt.be

_____

ABSTRACT

The model for the cumulative n^{th} citation distribution, as developed in [L. Egghe and I.K. Ravichandra Rao. Theory of first-citation distributions and applications. Mathematical and Computer Modelling 34, 81-90, 2001] is extended to the general source-item situation. This yields a time-dependent Lotka function based on a given (static) Lotka function (considered to be valid for time $t = ¥$). Based on this function, a time-dependent Lotkaian informetrics theory is then further developed by, e.g. deriving the corresponding time-dependent rank-frequency function.

---

[1] Permanent address
Key words: Lotka, Lotkaian informetrics, time-dependent, h-index, Hirsch, g-index

These tools are then used to calculate the dynamical (i.e. time-dependent) g-index (of Egghe) while also an earlier proved result on the time-dependent h-index (of Hirsch) is refound. It is proved that both indexes are concavely increasing to their steady state values for $t = \yen$.

I. Introduction

It is only now that the author of the present paper fully understands the general applicability of the method developed in [1] where one determines a formula for the cumulative n^{th} citation distribution. Let us, briefly, indicate its construction and then, further on in this introductory section, indicate its general informetrical application to an item-time dependent Lotkaian theory.

Note that in [2] and [3], a general model for the nth citation distribution based on stopping times has been developed (a model which is not used here).

The cumulative n^{th} citation distribution is the cumulative distribution of the times t at which documents in a (general) bibliography receive their n^{th} citation (distribution among the ever cited documents). Here $n \hat{I} \yen$ but in a continuous setting, $n \hat{I} \text{¡}^{+}$ (to be specified further). This cumulative distribution is constructed as follows. Let us consider a set of documents "existing" at time $t = 0$ (or, which we consider from time $t = 0$ onwards). The cumulative distribution of citations to these documents is denoted by $C(t)$, $t^3 \; 0$. So, for every $t > 0$, $C(t)$ denotes the fraction (among all citations up to time $t = \yen$) of the citations given to these documents (in total) at t. Besides $C(t)$ we also have the distribution of citations (at time $t = \yen$) among the documents. The fraction of documents (with respect to the ever cited documents) with j citations at $t = \yen$ is denoted $\varphi(j)$. In a continuous setting, j can take any value $j \hat{I} \; [1, +\yen[$ (being a density). In general one can work with the distribution $\varphi(j)$ as such, but in [1] (and also in this article) we will restrict ourselves to Lotkaian distributions $\varphi(j)$ on $[1, +\yen[$ of the type

$$\varphi(j) = \frac{\alpha - 1}{j^{\alpha}} \tag{1}$$

with $j \in [1,+\infty[$ and $\alpha > 1$. For more on Lotkaian informetrics, we refer the reader to [4]. The formula (1) expresses that $\varphi(j)$ is proportional to $\frac{1}{j^{\alpha}}$ (i.e. a decreasing power law) and the proportionality factor $\alpha - 1$ arrives based on the distributional requirement (as is readily checked):

$$\int_{1}^{\infty} \varphi(j)dj = 1 \qquad (2)$$

In [4], a more general Lotkaian theory is developed allowing for finite or infinite upper bounds for the variable j but in this paper (and in [1]), we limit ourselves to the (simplest) case $j \in [1,+\infty[$.

With these tools, the cumulative n^{th} citation distribution, denoted here by Ψ_n, is constructed as follows. For each document with j citations (density) in total at time $t = \infty$ we can assume that, if there is a fraction $C(t)$ of the total number of citations at time t, this document will have

$$jC(t) = n \qquad (3)$$

citations at time t. Of course, based on (3), documents will have n or more citations at time t if and only if their total number of citations is $j' \geq j$. Their fraction, by definition of φ, is given by

$$\int_{j}^{\infty} \varphi(A')dA' = \int_{j}^{\infty} \frac{\alpha - 1}{j'^{\alpha}}dj' = j^{1-\alpha} \qquad (4)$$

since $\alpha > 1$. This fraction is $\Psi_n(t)$ with n as in (3). Hence, by (3) and (4),

$$\Psi_n(t) = \left(\frac{n}{C(t)}\right)^{1-\alpha}$$

$$\Psi_n(t) = \left(\frac{C(t)}{n}\right)^{\alpha-1}$$

(5)

Formula (5) corresponds to formula (16) in [1] and represents the cumulative n^{th} citation distribution for $t \geq 0$ (with respect to the ever cited papers). Note that, for $t = 0$, $\Psi_n(0) = C(0) = 0$ and that, for $t = \yen$ (meaning in fact $\lim_{t \to \yen}$),

$$\Psi_n(\yen) = \left(\frac{C(\yen)}{n}\right)^{\alpha-1} = \frac{1}{n^{\alpha-1}}$$

(6)

since $C(t)$ is the cumulative distribution of the times of all citations, hence $C(\yen) = 1$. We see that $\Psi_n(\yen)$ is nothing else than the cumulative fraction of (ever cited – i.e. with at least one citation at $t = \yen$) documents that have n <u>or more</u> citations at $t = \yen$. One can, indeed, verify that

$$\Psi_n(\yen) = \int_n^{\yen} \varphi(A)dA$$

for all $n \in \yen$. Hence

$$\varphi(n) = -\frac{d\Psi_n(\yen)}{dn}$$

(7)

for all $n \in \yen$.

In the next section we will repeat the above argument in the general Lotkaian framework of sources and items (where "documents" above are replaced by general "sources" and "citations" are replaced by "items"). From the arguments we will derive the Lotka distribution that is valid at every $t \geq 0$. Also from this the rank-frequency function (of Zipfian type) will be derived. This gives a possible time dependent theory of Lotkaian informetrics, limited to the restriction (3).

The third section is devoted to applications of this time-dependent theory. It will be applied to the calculation of the, promising, h- and g-index, in a time dependent perspective, see [5],

[6], [7], [8], [9], [10], [11], [12] for a discussion on the h-index and [13], [14], [15] for the introduction of the g-index. Let us just define the h- and g-index. Suppose an author's publications are ranked in decreasing order of the number of citations received. Then h is the largest rank such that all publications on ranks 1,…,h have h or more citations. Note that, once an article is in the "leading" group, it does not influence a possible further increase of h, even when these articles receive many more citations. Since we feel that an index for this purpose (being the measuring of the overall citation performance of an author) should be influenced by the top-scoring articles, we modified, in [13], [14], [15] the h-index as follows. Note that it follows from the definition of the h-index that the top h articles receive, together, h^2 or more citations. Therefore the following definition of the g-index has been given.

Suppose again that an author's publications are ranked in decreasing order of the number of citations received. Then g is the largest rank such that all publications on rank 1,…,g have, together, g^2 or more citations. Evidently, $g \geq h$ and now a possible increase of g is also dependent on the future performance (in the sense of citations received) of the top articles.

From [11], [13], [14], [15], it is clear that the h- and g-index can be calculated in any (Lotkaian) informetric system. Hence, since this article presents a time-dependent Lotkaian informetrics theory, we are able, in this article, to present time-dependent formulae for the h- and g-index. Here we refind the formula in [16] for the time-dependent h-index but the formula for the g-index, presented in Section 3, is new. We show that, in both cases, the time-dependent h- and g-index is a concavely increasing function.

II. A time-dependent Lotkaian informetrics theory

Let us have a general information production process (IPP) of sources and items (see [4]). Let the Lotkaian distribution be given as in (1):

$$\varphi(j) = \frac{\alpha - 1}{j^{\alpha}}$$

This situation is considered to be valid at $t = \yen$ and we will deduce, from this, the time evolution of such an IPP. We, therefore need a distribution which describes the increase of the items over time t. Let $G(t)$ denote this cumulative distribution of the growth of the number of items (since we generalized our terminology from citations to general items we will use the notation $G(t)$ for the cumulative fraction of items (growth) at time t).

In formula (3) we used the notation n since n appeared in [1] for the cumulative n^{th} citation distribution. In this general setting we will replace n by k since it is more in line with the notation j for item densities at $t = \yen$. So equation (3) now reads

$$jG(t) = k \tag{8}$$

meaning that a source with j items (density) at $t = \yen$ will have k items (density) at t. Of course, sources will have k or more items if and only if their total number of items is $j' \geq j$. Their fraction, by definition of φ is given by

$$\int_{j}^{\yen} \varphi(j')dj' = j^{1-\alpha} \tag{9}$$

(as in Section I). This fraction is hence the cumulative fraction $\Phi(k,t)$ of sources that have, at t, k or more items. Hence, by (8) and (9) we have

$$\Phi(k,t) = \left(\frac{G(t)}{k} \right)^{\alpha-1} \tag{10}$$

for $k \geq G(t)$. ($\Phi(k,t)$ is the same as $\Psi_{k}(t)$ in (5) but now we consider k as a variable which we did not in the derivation of the cumulative n^{th} (or k^{th}) citation distribution.)

Denoting by $\varphi(k,t)$ the time-dependent Lotka function we have

$$\Phi(k,t) = \int_{k}^{\yen} \varphi(k',t)dk' \tag{11}$$

Hence

$$\varphi(k,t) = -\frac{d\Phi}{dk}(k,t)$$

$$\varphi(k,t) = (\alpha - 1)\frac{G(t)^{\alpha-1}}{k^{\alpha}} \tag{12}$$

for $k \geq G(t)$ (by (8) and since $j \geq 1$). In other words, by (1),

$$\varphi(k,t) = G(t)^{\alpha-1}\varphi(k) \tag{13}$$

for $k \geq G(t)$ (hence, here, the function (1) is adopted from $G(t)$ onwards). Hence, by (12), up to a constant, $\varphi(k,t)$ is $\varphi(k)$ but remember that $k \geq G(t)$ which is < 1 while $\varphi(j)$ in (1) only is defined from $j \geq 1$ onwards. Nevertheless we can say that the decreasing power law (1) with exponent α in (1) is maintained in (12), from $k \geq G(t)$ onwards.

To be able to calculate actual number of sources we define

$$f(k,t) = T\varphi(k,t) \tag{14}$$

$$f(k,t) = TG(t)^{\alpha-1}\varphi(k) \tag{15}$$

$$f(k,t) = G(t)^{\alpha-1}f(k) \tag{16}$$

where $f(k)$ is defined as $T\varphi(k)$ and where T denotes the total number of sources at time $t = \yen$. Hence, for every $k \geq G(t)$,

$$\int_{k}^{\yen} f(k',t)dk' \tag{17}$$

is the total number of sources with k or more items at time t. This will be used further on in the determination of the rank-frequency function $\gamma(r,t)$, associated with the size-frequency function $f(k,t)$ at time t.

Example

A classical example is one where $G(t)$ is the cumulative exponential distribution, expressing concave growth. Hence $G(t)$ is of the form

$$G(t) = 1 - a^t \tag{18}$$

where $0 < a < 1$. Note that $\lim_{t \to \infty} G(t) = 1$, as required. Then the time-dependent Lotkaian distribution (13) reads, using also (1):

$$\varphi(k,t) = \left(1 - a^t\right)^{\alpha - 1} \frac{\alpha - 1}{j^{\alpha}} \tag{19}$$

and similar for (19).

Let us first calculate $T(t)$, $A(t)$, the total number of sources and items respectively at time t. We have, evidently (since $k \geq G(t)$)

$$T(t) = \int_{G(t)}^{\infty} f(k,t)dk$$

$$T(t) = TG(t)^{\alpha - 1} \int_{G(t)}^{\infty} \varphi(k)dk$$

$$T(t) = T \tag{20}$$

(using (1) and (15)).

$$A(t) = \int_{G(t)}^{\infty} kf(k,t)dk$$

$$A(t) = TG(t)^{\alpha-1} \int_{G(t)}^{\yen} k\varphi(k)dk$$

$$= TG(t)^{\alpha-1} \frac{\alpha-1}{\alpha-2} G(t)^{2-\alpha} \tag{21}$$

But since

$$\mu = \frac{A}{T} = \int_1^{\yen} k\varphi(k)dk = \frac{\alpha-1}{\alpha-2} \tag{22}$$

, being the average number of items per source at $t = \yen$, it follows from (21) and (22) that

$$A(t) = AG(t) \tag{23}$$

It follows from (20) and (23) that the average number of items per source at t, $\mu(t)$ is given by

$$\mu(t) = \frac{A(t)}{T(t)} = \frac{AG(t)}{T} = \mu G(t) \tag{24}$$

Results (20), (23) and (24) are logical results given (8) (and where there is no change in source rankings, hence also not in the total number of sources).

From (20) it also follows that, for the rank-frequency function $\gamma(r,t)$, we can take $r \hat{I} [0,T]$ as is the case for $t = \yen$.

By (17), the defining relation for the rank-frequency function $\gamma(r,t)$, (by (10) and since the number of sources do not change) is

$$r = \gamma^{-1}(k,t) = \int_k^{\yen} f(k',t)dk'$$

where γ^{-1} refers to the inverse of the function $r \circledR \gamma(r,t)$ and where $r \hat{I} [0,T]$. Hence, by (1) and (15) we have

$$r = \gamma^{-1}(k,t) = TG(t)^{\alpha-1} \grave{O}_k^{\yen} \frac{\alpha-1}{k^{'\alpha}} dk' \tag{25}$$

Denote $C = T.(\alpha-1)$ being the constant in the numerator of f:

$$f(j) = \frac{T(\alpha-1)}{j^{\alpha}} = \frac{C}{j^{\alpha}}$$

(25) gives

$$r = \frac{CG(t)^{\alpha-1}}{(\alpha-1)k^{\alpha-1}}$$

So

$$k^{\alpha-1} = \frac{CG(t)^{\alpha-1}}{(\alpha-1)r}$$

and, finally, by definition of γ, for $r \hat{I} [0,T]$,

$$k = \gamma(r,t) = \frac{C^{\frac{1}{\alpha-1}}}{((\alpha-1)r)^{\frac{1}{\alpha-1}}} G(t) \tag{26}$$

Now note (cf. Exercise II.2.2.6, p. 134 in [4] or see the Appendix in [11] for a proof) that, for $r \hat{I} [0,T]$,

$$\gamma(r) = \frac{C^{\frac{1}{\alpha-1}}}{((\alpha-1)r)^{\frac{1}{\alpha-1}}} \tag{27}$$

is the rank-frequency function at $t = ¥$ (a direct proof can be given as above for $\gamma(r,t)$ but now using the function $f(k)$ at $t = ¥$). Hence it follows from (26) and (27) that

$$\gamma(r,t) = \gamma(r)G(t) \qquad (28)$$

a logical result.

This ends our basic study of time-dependent Lotkaian informetrics. We will now apply this theory to the calculation of the time-dependent h- and g-indexes.

III. The time-dependent h- and g-index

III.1 The time-dependent h-index

The time-dependent h-index is mathematically defined as:

$$\grave{\mathrm{O}}_{h(t)}^{¥} f(k,t)dk = h(t) \qquad (29)$$

Alternatively one can define $h(t)$ by (see [11])

$$\gamma\big(h(t),t\big) = h(t) \qquad (30)$$

The reader will notice that these definitions are exact model theoretic formulations of the (time-dependent) h-index as given in the introduction.

From (29), (1) and (15) we have

$$TG(t)^{\alpha-1} \grave{\mathrm{O}}_{h(t)}^{¥} \frac{\alpha-1}{k^{\alpha}} dk = h(t)$$

whence

$$TG(t)^{\alpha-1}h^{1-\alpha} = h(t)$$

hence

$$h(t) = G(t)^{\frac{\alpha-1}{\alpha}} T^{\frac{1}{\alpha}} \qquad (31)$$

Note that

$$\lim_{t \to \yen} h(t) = T^{\frac{1}{\alpha}} = h \qquad (32)$$

, the h-index for $t = \yen$ (cf. [11]). So we refind the result (31) or (33) as proved in [16]:

$$h(t) = G(t)^{\frac{\alpha-1}{\alpha}} h \qquad (33)$$

but there one was only dealing with the "citation" case for the items.

A second proof of these results can be given via (30):

$$\gamma(h(t),t) = \frac{C^{\frac{1}{\alpha-1}}}{((\alpha-1)h(t))^{\frac{1}{\alpha-1}}} G(t) = h(t)$$

whence (31), using that $C = T(\alpha - 1)$.

III.2 The time-dependent g-index

The time-dependent g-index is mathematically defined as: if

$$\int_{k}^{\yen} k' f(k',t)dk' = g(t)^{2} \qquad (34)$$

then

$$g(t) = \gamma^{-1}(k,t) \tag{35}$$

Alternatively one can replace these conditions by (cf. [14])

$$\int_0^{g(t)} \gamma(r,t)dr = g(t)^2 \tag{36}$$

Again the reader can verify that this is the mathematically exact formulation of the g-index, discussed in the introduction (note we must add the condition "if $g(t) £ T$; otherwise we take $g(t) = T$" – see [14]).

Now (36) and (26) yield

$$\int_0^{g(t)} \frac{C^{\frac{1}{\alpha-1}}}{((\alpha-1)r)^{\frac{1}{\alpha-1}}} G(t)dr = g(t)^2$$

$$\left(\frac{C}{\alpha-1}\right)^{\frac{1}{\alpha-1}} G(t)\int_0^{g(t)} r^{-\frac{1}{\alpha-1}}dr = g(t)^2$$

$$\left(\frac{C}{\alpha-1}\right)^{\frac{1}{\alpha-1}} G(t)\frac{\alpha-1}{\alpha-2} g(t)^{\frac{\alpha-2}{\alpha-1}} = g(t)^2$$

whence

$$g(t)^{\frac{\alpha}{\alpha-1}} = T^{\frac{1}{\alpha-1}} G(t)\frac{\alpha-1}{\alpha-2}$$

using that $C = T.(\alpha-1)$. Hence we proved the following formula for $g(t)$:

$$g(t) = G(t)^{\frac{\alpha-1}{\alpha}} T^{\frac{1}{\alpha}} \left(\frac{\alpha-1}{\alpha-2}\right)^{\frac{\alpha-1}{\alpha}} \tag{37}$$

Note that

$$\lim_{t \to \infty} g(t) = T^{\frac{1}{\alpha}} \left(\frac{\alpha - 1}{\alpha - 2} \right)^{\frac{\alpha-1}{\alpha}} = g \tag{38}$$

the g-index for $t = \infty$ (cf. [14]). So we have also

$$g(t) = G(t)^{\frac{\alpha-1}{\alpha}} g \tag{39}$$

The same result follows when (34) and (35) are used to calculate $g(t)$.

From (33) and (39) it also follows that

$$\frac{g(t)}{h(t)} = \frac{g}{h} \tag{40}$$

Or otherwise stated $h(t)$ and $g(t)$ grow at the same pace.

III.3 Shape of the functions h(t) and g(t)

From (33) and (38) it is clear that we know the shape of $h(t)$ and $g(t)$ if we know the shape

of the function $t \to G(t)^{\frac{\alpha-1}{\alpha}} =: B(t)$. We have

$$B'(t) = \frac{\alpha - 1}{\alpha} G(t)^{-\frac{1}{\alpha}} G'(t)$$

Since we have that G strictly increases we have that $B'(t) > 0$ for all $t > 0$, hence B strictly

increases. Furthermore

$$B''(t) = \frac{\alpha - 1}{\alpha} \left(-\frac{1}{\alpha} G(t)^{-\frac{1}{\alpha}-1} G'^2(t) + \frac{\alpha - 1}{\alpha} G(t)^{-\frac{1}{\alpha}} G''(t) \right)$$

Since G is concavely increasing (to the asymptote at ordinate 1) we find that $B(t)$ is concavely increasing (hence the same is true for $h(t)$ and $g(t)$). Furthermore

$$\lim_{\substack{t \to 0 \\ >}} B(t) = 0$$

$$\lim_{t \to \yen} B(t) = 1$$

$$\lim_{\substack{t \to 0 \\ >}} B'(t) = +\yen$$

(since $G(0) = 0$ and $\alpha > 1$ and $G'(0) > 0$)

$$\lim_{t \to \yen} B'(t) = 0$$

(since the same is true for G'). Using (32) and (38) we hence have the shape of $h(t)$ and $g(t)$ as in Fig.1.
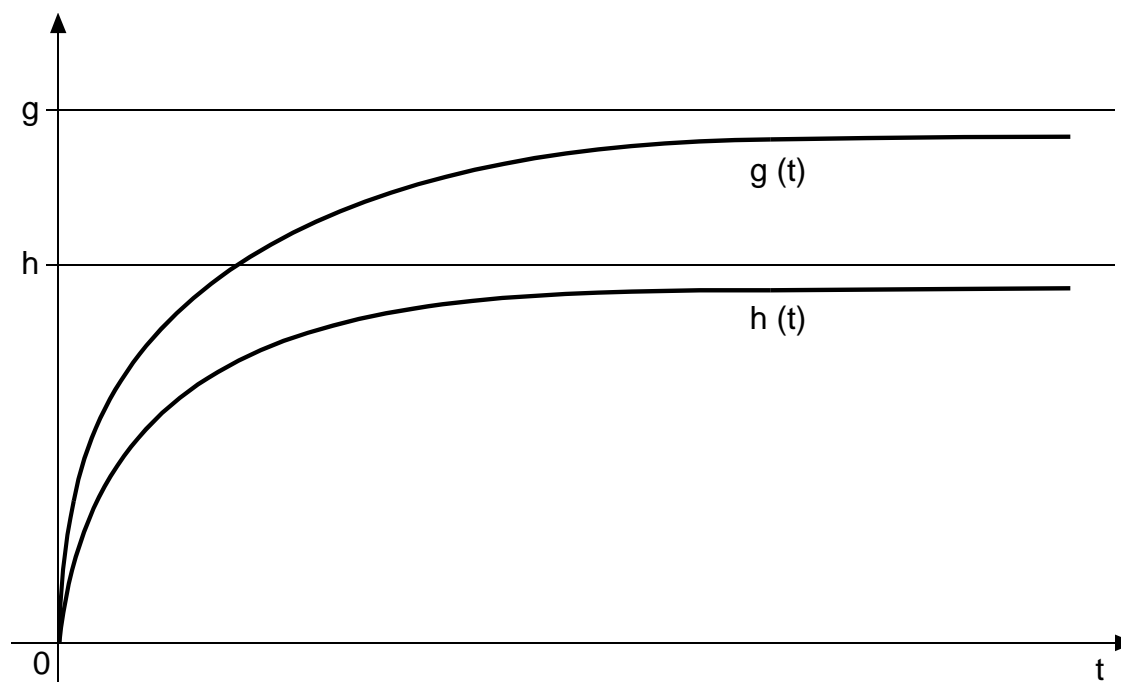

Fig. 1. The functions h(t) and g(t)

IV. Conclusions and further research

The theory of modelling the cumulative <u>first</u> citation distribution, i.e. the cumulative distribution of times at which papers receive their first citation was presented originally in [17] and further generalized in [1]. In the latter paper the only remark how to extend this theory to the modelling of the cumulative n^{th} citation distribution is limited to a half page (lower half of p.84 in [1]). Here it is even indicated that it is a "less important" generalization of the model of the cumulative first citation distribution!

At that time, the cumulative n^{th} citation distribution was only considered as a function of time t, with n fixed. The paper [16] taught us that this theory could be applied to the calculation of the time-dependent h-index. It then came apparent that the cumulative n^{th} citation distribution could also be considered as a function of n and t where, for $t = \yen$, the classical Lotkaian model is given. This then gave the idea of generalizing the theory of the cumulative n^{th} citation distribution to a theory of time-dependent Lotkaian informetrics: documents and citations are generalized to sources and items and the model of the cumulative n^{th} citation distribution is further elaborated to a model for time-dependent size- and rank-frequency functions: this is done in this paper.

Having at our disposal such a time-dependent Lotkaian theory we are, of course, also able to calculate the h-index $h(t)$ and the g-index $g(t)$ in function of time and this in the general source-item framework. We show that, both $h(t)$ and $g(t)$, are concavely increasing functions with a horizontal asymptote at the values h (respectively g) of the h- and g-index for time $t = \yen$.

The present model lacks one more important generalization: the present model is constrained to the sources to be constant, i.e. they are there from $t = 0$ onwards (albeit with less items, the smaller t is). In a forthcoming paper we will investigate how to incorporate another growth function into this model so that the number of sources can also grow in time. We are working on such a general informetric model that incorporates growth in sources as well as in items. Such a theory can then be further applied to the more general evolution of the h- and g-index in such systems.

<u>References</u>

[1]   L. Egghe and I.K. Ravichandra Rao (2001). Theory of first-citation distributions and applications. Mathematical and Computer Modelling 34, 81-90.

[2]   W. Glänzel (1992). On some stopping times on citation processes. From theory to indicators. Information Processing and Management 28(1), 53-60.

[3]   W. Glänzel and U. Schoepflin (1995). A bibliometric study on aging and reception processes of scientific literature. Journal of Information Science 21(1), 37-53.

[4]   L. Egghe (2005). Power Laws in the Informetric Production Process: Lotkaian Informetrics. Elsevier, Oxford, UK.

[5]   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, 16569-16572.

[6]   P. Ball (2005). Index aims for fair ranking of scientists. Nature 436, 900.

[7]   L. Bornmann and H.-D. Daniel (2005). Does the h-index for ranking of scientists really work? Scientometrics 65, 391-392.

[8]   T. Braun, W. Glänzel and A. Schubert (2005). A Hirsch-type index for journals. The Scientist 19(22), 8.

[9]   W. Glänzel (2006a). On the opportunities and limitations of the H-index. (In Chinese). Science Focus, 1(1), 10-11.

[10]   W. Glänzel (2006b). On the H-index – a mathematical approach to a new measure of publication activity and citation impact. Scientometrics 67(2), 315-321.

[11]   L. Egghe and R. Rousseau (2006). An informetric model for the Hirsch index. Scientometrics, to appear.

[12]   A.F.J. Van Raan (2005). Comparison of the Hirsch-index with standard bibliometric indicators and with peer judgement for 147 chemistry research groups. Scientometrics 67(3), to appear. http://arxiv.org/abs/physics/0511206.

[13]   L. Egghe (2006a). How to improve the h-index. The Scientist, 20(3), 14, March 2006.

[14]   L. Egghe (2006b). Theory and practise of the g-index. Scientometrics, to appear.

[15]   L. Egghe (2006c). An improvement of the h-index: the g-index. ISSI Newsletter 2(1), 8-9.

[16]   L. Egghe (2006d). Dynamic h-index: the Hirsch index in function of time. Journal of the American Society for Information Science and Technology, to appear.

[17]   L. Egghe (2000). A heuristic study of the first-citation distribution. Scientometrics 48(3), 345-359.