Available online at www.sciencedirect.com
ScienceDirect
Journal of Informetrics 1 (2007) 92–102
Journal of INFORMETRICS
An International Journal
www.elsevier.com/locate/joi

Characteristic scores and scales
A bibliometric analysis of subject characteristics based on long-term citation observation☆

Wolfgang Glänzel[a,b,*]

[a] *Steunpunt O&O Statistieken, K.U. Leuven, Dekenstraat 2, B-3000 Leuven, Belgium*
[b] *IRPS, Hungarian Academy of Sciences, Budapest, Hungary*

Received 3 July 2006; received in revised form 5 October 2006; accepted 19 October 2006

Abstract

In an earlier paper by Glänzel and Schubert [Glänzel, W., & Schubert, A. (1988a). Characteristic scores and scales in assessing citation impact. *Journal of Information Science*, *14*(2), 123–127; Glänzel, W., & Schubert, A. (1988b). Theoretical and empirical studies of the tail of scientometric distributions. In L. Egghe, & R. Rousseau (Eds.), *Informetrics*: *Vols. 87/88*, (pp. 75–83). Elsevier Science Publisher B.V.], a method for classifying ranked observations into self-adjusting categories was developed. This parameter-free method, which was called method of characteristic scores and scales, is independent of any particular bibliometric law. The objective of the present study is twofold. In the theoretical part, the analysis of its properties for the general form of the Pareto distribution will be extended and deepened; in the empirical part the citation history of individual scientific disciplines will be studied. The chosen citation window of 21 years makes it possible to analyse dynamic aspects of the method, and proves sufficiently large to also obtain stable patterns for each of the disciplines. The theoretical findings are supplemented by regularities derived from the long-term observations.

Keywords: Citation analysis; Long-term citation impact; Disciplinary citation impact; Pareto distribution; Extreme values; Truncated moments

1. Introduction

The classical studies of long-term citation impact by Abt (1981) and Garfield (1996, 1998a, 1998b) have shown that long observation periods are indispensable for obtaining reliable and stable results on citation processes. From the methodological viewpoint, such analyses are important in studying disciplinary citation impact, ageing-related issues, first-citation distributions (see, for instance, Egghe & Rao, 2001; Glänzel & Schoepflin, 1995; Rousseau, 1994) and citation-succession processes (see Glänzel, 1992; Glänzel & Schoepflin, 1995), the phenomenon of delayed recognition (e.g., Garfield, 1980; Glänzel & Garfield, 2004; Glänzel, Schlemmer, & Thijs, 2003; van Raan, 2004) and in identifying highly cited papers. Informetric long-term studies reveal regularities and provide tools that are nevertheless important for scientometric applications, too. The predictive power of models for citation processes (Glänzel & Schubert, 1995),

☆ Present study is partially based on a talk delivered at the "10th Nordic Workshop on Bibliometrics, Informetrics and Research Policy" (Glänzel, 2005).

* Present address: Steunpunt O&O Statistieken, K.U. Leuven, Dekenstraat 2, B-3000 Leuven, Belgium.
E-mail address: Wolfgang.Glanzel@econ.kuleuven.be.

1751-1577/$ – see front matter
doi:10.1016/j.joi.2006.10.001

the methodological foundation of finding optimum citation windows for evaluative purposes and the determination of subject-specific scores are only some of those applications.

In the present paper, we will use a method for classifying ranked observations into self-adjusting categories developed by Glänzel and Schubert (1988a, 1988b). The method, which is based on the analysis of rank-specific subcategories of ranked observations, was called characteristic scaling. Within the framework of this model an optionally refinable scale with variable scores is used for the characterising eminence of citation impact. It can be shown that if the grouping derived by our procedure is complete for the ranked observations it has several interesting properties, the most important of which is the grouping's independence of the common scale. Such scale- and parameter-independent classification methods are very important in scientometric practice since citation distributions are highly sensitive to subject particularities (see, for instance, Glänzel & Moed, 2002). Moreover, instead of the usual comparison of statistical functions like means, medians or percentages of a sub-sample with the corresponding values of the complete sample or of the population, this method directly allows the benchmarking of the sub-sample through the comparison of its class properties with that of the population. This, of course, provides a more complex approach than the usual comparison of "averages".

In the present paper we will first extend and generalise the investigation of the theoretical properties of the method under the condition that the underlying citation distribution is a *Pareto distribution of the second kind* at any time beginning with the publication year. In order to obtain stable results and to be able to analyse the dynamics of the *characteristic scores and scales* the study is based on a 21-year citation window. We will also demonstrate that the parameter α (and its transformation $q=(1-1/\alpha)^{\alpha}$) derived from our model describes important subject-specific characteristics of citation distributions. Although the method of *characteristic scores and scales* itself represents a parameter-free approach to citation-impact classification, it provides the characteristic parameter of the underlying distribution. Furthermore, the analysis of the citation process gives empirical evidence that this parameter is time-invariant.

2. The theory of 'characteristic scores and scales'

In what follows, we briefly summarise the definition of the *characteristic scores and scales* according to Glänzel & Schubert (1988a, 1988b). Consider a set of n papers published in a given subject field. The observed citations $\{X_i\}_{i=1}^{n}$ received by each paper in a given time period are then ranked in descending order $X_1^* \geq X_2^* \geq \cdots \geq X_n^*$, where X_1^* denotes the citation rate of most frequently cited paper in the set and X_n^* consequently the number of citations the least cited paper has received.

2.1. Definition

In order to develop a method for subdividing the sample into classes we define appropriate thresholds based on the following recursion. First put $\beta_0=0$ and $\nu_0=n$. β_1 is then defined as the sample mean:

$$\beta_1=\sum_{i=1}^{n}\frac{X_i}{n}=\sum_{i=1}^{n}\frac{X_i^*}{\nu_0} \tag{1}$$

The value ν_1 is defined by the following inequality:

$$X_{\nu_1}^* \geq \beta_1 \quad \text{and} \quad X_{\nu_1+1}^* < \beta_1 \tag{2}$$

This procedure is repeated recurrently, particularly:

$$\beta_k=\sum_{i=1}^{\nu_{k-1}}\frac{X_i^*}{\nu_{k-1}} \tag{3}$$

and ν_k is chosen so that:

$$X_{\nu_k}^* \geq \beta_k \quad \text{and} \quad X_{\nu_k+1}^* < \beta_k, \quad k \geq 2 \tag{4}$$

The properties $\beta_0 \leq \beta_1 \leq \ldots$ and $\nu_0 \geq \nu_1 \geq \ldots$ are obvious from the definition. Obviously, the procedure comes to an end if $\nu_k=1$ for some $k>0$ is reached. The kth class is defined by the pair of threshold values $[\beta_{k-1}, \beta_k)$ and the

number of papers belonging to this class amounts to $\nu_{k-1} - \nu_k$. From the definition defining procedure it follows that each given publication set determines its own groups.

2.2. Properties

Though no underlying rule is necessary for arranging the sample, important properties of the threshold values as well as of the size of the classes determined through these thresholds can be derived for special citation distributions. Since citation distributions are integer valued, the unique determination of the theoretical values of the characteristic thresholds introduced above would always result in an integer approximation. Therefore, we have used an approximation based on a continuous distribution model in an earlier study (Glänzel, Telcs, & Schubert, 1984). In particular, we have found an important basic property in the case of Paretian distributions. The citation rate X of a paper has a Paretian distribution if

$$G(x) := 1 - F(x) = P(X \geq x) \approx c(N + x)^{-\alpha} \tag{5}$$

for large $x > 0$ and some positive value c, where N and α are positive real parameters and F denotes the distribution function of the random variable X. For $\alpha > 1$ the citation rate X has a finite expectation so that one can define the following conditional expectation:

$$b_k = \begin{cases} 0 & k = 0 \\ \dfrac{\sum_{i \geq b_{k-1}} i P(X = i)}{\sum_{i \geq b_{k-1}} P(X = i)} & k > 0 \end{cases} \tag{6}$$

According to the characterisation theorem for Paretian distributions by Glänzel et al. (1984) the conditional expectation satisfies the condition $b_k = E(X|X \geq b_{k-1}) \sim \{\alpha/(\alpha - 1)\} b_{k-1} + b_1$, which, in turn, results by recursion in the following property:

$$b_k \approx b_1 \sum_{i=1}^{k-1} \left\{ \frac{\alpha}{\alpha - 1} \right\}^i \tag{7}$$

with b_1 being the expected value and α the characteristic parameter of the Pareto-approximation. Since the normalised scores $b_k^* = b_k / b_1$ depend only on the characteristic parameter we have called the method *characteristic scaling* and the values $b_k^{(*)}$ as well as the corresponding statistics $\beta_k (\beta_k / \beta_1)$ themselves *characteristic scores*. If the expectation of the citation distribution is finite, that is, if $\alpha > 1$ (or, in the finite case, if $\alpha < 0$) the values β_k defined above are estimators of the theoretical values b_k.

In the present study we will go a step farther. First we modify the model as follows. Instead of the Pareto-approximation for large x (see Eq. (1)) we will use the Pareto distribution as the underlying model for received citation rates. This does, because of the assumed continuity of the random variable, is not quite in keeping with the discreteness of citation rates, but can readily be used, for instance, as a continuous approximation of its discrete analogue, namely the Waring distribution (cf. Glänzel et al., 1984).

The general form of Pareto distribution, also referred to as *Pareto distribution of the second kind* or *Lomax* distribution, can be obtained from the infinite beta distribution if one of the parameters is chosen 1 (see, e.g., Johnson, Kotz, & Balakrishnan, 1994). In particular, we say that the non-negative random variable X has a Pareto distribution (of the second kind) if:

$$G(x) = P(X \geq x) = \frac{N^\alpha}{(N + x)^\alpha}, \quad \text{for all } x \geq 0 \tag{8}$$

where N and α are positive real parameters. Alternatively, the parameter transformation $a = \alpha/(\alpha - 1)$ is used as well.

In a second step, instead of the ranked sample elements $X^*_{\nu_k}$ the corresponding theoretical values, namely Gumbel's so-called characteristic kth extreme values (Gumbel, 1958) will be used. The authors have shown that for large samples ($n \gg 1$) and relatively small $k \ll n$ a small correction of these extreme values results in modified Gumbel's extreme values according to Glänzel and Schubert (1988a, 1988b) which actually form the median of the corresponding order statistic. However, for reason of simplicity of calculation we will not apply this correction. If the distribution function F of the random variable X is absolutely continuous and strictly monotonous (as, for instance, the Pareto distribution) we

can define the characteristic kth extreme value as $u_k := G^{-1}(k/n)$. Assuming these properties of the distribution function we can then define the theoretical group size m_k through the characteristic extreme values and the characteristic thresholds b_k as $u_n := b_k$. Thus, we can consider the statistics β_k and ν_k estimators of the corresponding theoretical values b_k and m_k. The following property is obvious:

$$G(b_k) = G(u_{m_k}) = \frac{m_k}{n} \tag{9}$$

In other words, the theoretical class sizes $m_{k-1} - m_k$ $(k > 0)$ can be derived from the distribution function as follows:

$$m_k - m_{k-1} = n\{G(b_k) - G(b_{k-1})\} \tag{10}$$

If $\alpha > 1$ the expected value of the Pareto distribution is finite and can be expressed by $b_1 = N/(\alpha - 1)$, or alternatively by $b_1 = N(a-1)$ using $a = \alpha/(\alpha - 1) > 1$. Hence, and from the characterisation theorem for Pearson-type distributions (Glänzel et al., 1984) we obtain the characteristic thresholds b_k by the following recursion:

$$b_k = E(X|X \geq b_{k-1}) = ab_{k-1} + b_1 = N(a-1)\left(\sum_{i=0}^{k-1} a^i\right) = N(a^k - 1) \tag{11}$$

Hence, we have

$$\frac{b_k}{b_{k-1}} = \frac{a^k - 1}{a^{k-1} - 1} = a + \frac{1}{\sum_{i=0}^{k-2} a^i} \tag{12}$$

The following three properties are obvious:

(i) $b_2^* = \frac{b_2}{b_1} = a + 1$
(ii) $\frac{b_k}{b_{k-1}} \in \left(a,\ \frac{a+1}{k-1}\right)$ if $k > 1$
(iii) $\frac{b_k}{b_{k-1}} \sim a$ if $k \gg 1$

For the class size $(m_{k-1} - m_k)$ we obtain the following important properties through the characterisation theorem and the definition of Gumbel's extreme values:

$$G(u_{m_k}) = G(b_k) = N^\alpha (N + N(a^k - 1))^{-\alpha} = a^{-k\alpha} = \left[\frac{\alpha - 1}{\alpha}\right]^{k\alpha} = \left(\frac{1 - 1}{\alpha}\right)^{k\alpha} = q^k \tag{13}$$

where $q := (1 - 1/\alpha)^\alpha$. Consequently, we have

$$\frac{m_k}{m_{k-1}} = \frac{m^{k-1} - m_k}{m^{k-2} - m^{k-1}} = q = \left(\frac{1 - 1}{\alpha}\right)^\alpha \quad \text{for all } k > 1 \tag{14}$$

The $\alpha - q$ relationship is presented in Table 1. The most relevant α range in bibliometrics is shaded. $\alpha = 1$ corresponds to a Lotka-type distribution.

In addition to the Pareto distribution we also have a look at one of the important limiting cases, namely the *exponential distribution*. In particular, if $N, \alpha \to \infty$ and $N/\alpha \to \lambda$ for some finite real $\lambda > 0$ we obtain $G(x) = N^\alpha/(N + x)^\alpha \to \mathrm{e}^{-x/\lambda}$. Furthermore, $\alpha \to \infty$ implies $q = (1 - 1/\alpha)^\alpha \to \mathrm{e}^{-1}$ (cf. Table 1). Analogously to Eqs. (11)–(14) we obtain the properties of the characteristic scores and classes in the case of the exponential distribution as follows:

$$b_k = E(X|X \geq b_{k-1}) = kb_1 = k\lambda \quad \text{and} \quad b_k^* = \frac{b_k}{b_1} = k \tag{15}$$

$$\frac{b_k}{b_{k-1}} = \frac{k}{k-1} \sim 1 \quad \text{if } k \gg 1 \tag{16}$$

$$G(u_{m_k}) = G(b_k) = \mathrm{e}^{(-\lambda k/\lambda)} = \mathrm{e}^{-k} = q^k, \quad \text{where } q = \mathrm{e}^{-1} \tag{17}$$

$$\frac{m_k}{m_{k-1}} = \frac{m_{k-1} - m_k}{m_{k-2} - m_{k-1}} = q = \mathrm{e}^{-1} \quad \text{for all } k > 1 \tag{18}$$

Table 1
q as a function of α. The values most frequently observed in bibliometrics are shaded

α	q	α	q	α	q	α	q	α	q
1.00	0.0000	3.00	0.2963	5.00	0.3277	10	0.3487	30	0.3617
1.10	0.0715	3.10	0.2990	5.25	0.3298	11	0.3505	35	0.3626
1.20	0.1165	3.20	0.3015	5.50	0.3316	12	0.3520	40	0.3632
1.30	0.1486	3.30	0.3038	5.75	0.3333	13	0.3533	45	0.3638
1.40	0.1731	3.40	0.3060	6.00	0.3349	14	0.3543	50	0.3642
1.50	0.1925	3.50	0.3080	6.25	0.3363	15	0.3553	55	0.3645
1.60	0.2082	3.60	0.3099	6.50	0.3376	16	0.3561	60	0.3648
1.70	0.2213	3.70	0.3117	6.75	0.3388	17	0.3568	65	0.3650
1.80	0.2323	3.80	0.3133	7.00	0.3399	18	0.3574	70	0.3652
1.90	0.2418	3.90	0.3149	7.25	0.3409	19	0.3580	75	0.3654
2.00	0.2500	4.00	0.3164	7.50	0.3419	20	0.3585	80	0.3656
2.10	0.2572	4.10	0.3178	7.75	0.3428	21	0.3589	85	0.3657
2.20	0.2636	4.20	0.3191	8.00	0.3436	22	0.3594	90	0.3658
2.30	0.2692	4.30	0.3204	8.25	0.3444	23	0.3597	95	0.3659
2.40	0.2743	4.40	0.3216	8.50	0.3451	24	0.3601	100	0.3660
2.50	0.2789	4.50	0.3227	8.75	0.3458	25	0.3604	150	0.3666
2.60	0.2830	4.60	0.3238	9.00	0.3464	26	0.3607	200	0.3670
2.70	0.2868	4.70	0.3249	9.25	0.3470	27	0.3610	250	0.3671
2.80	0.2902	4.80	0.3258	9.50	0.3476	28	0.3612	300	0.3673
2.90	0.2934	4.90	0.3268	9.75	0.3482	29	0.3614	∞	0.3679

The first important property of the Pareto distribution (and the exponential distribution as a limiting case too) is the independence of the group size of the parameter N (or λ in the case of the exponential distribution) while in both cases the characteristic scores are proportional to these parameters. Furthermore, the ratio of the size of subsequent classes is constant according to Eqs. (14) and (18). The ratios of class sizes ($m_{k-1} - m_k$) with respect to the lowest class ($m_0 - m_1$) consequently form a geometric series $q{:}q^2{:},\ldots,{:}q^k$, which might be considered an *inverse Bradford law*. Unlike in the original Bradford law, the number of produced items in the zones or classes, that is the number of citations, is not constant. This law will form the theoretical base for the following empirical study.

3. Characteristic scores and scales in practice

The above-mentioned properties of characteristic scores and scales and of the classes defined by them hold for continuous Pareto distributions. Informetric distributions such as publication activity and citation impact are however discrete integer-valued distributions resulting from processes with continuous or discrete time parameter. In order to obtain robust and interpretable models that can also describe the changing shape of the distribution as time elapses, for instance, simple birth processes (Glänzel & Schoepflin, 1994) or mixtures of a Poisson process with appropriate continuous distributions such as the Gamma distribution (Burrell, 1990; Burrell & Cane, 1982) have been assumed. Both models result in negative binomial processes that can describe important features such as changing citation impact in time or the ageing of information but fail to model the tail properties of informetric distributions (Glänzel & Schubert, 1988a, 1988b). The assumption of generalised Yule processes (Schubert & Glänzel, 1983) or further mixture, e.g., with a Pareto distribution result in processes having a (generalised) Waring or Irwin distribution (e.g., Karlis & Xekalaki, 2005; Xekalaki, 1983) at any time. These models describe the tail properties of informetric distribution in an appropriate manner, and their tail can be approximated by the Pareto distribution. Therefore, one could expect that the above model fits sufficiently well empirical data. In what follows, we will prepare the scores and scale to define an appropriate set of classes (zones) to characterise citation distributions and to set thresholds to distinguish poor, fair, remarkable and outstanding citation impact. It is a consequence from their definition, characteristic scores and scales are particularly suited to assessing excellent and outstanding citation rates. As has shown in the previous section, the

number of scores is finites since the procedure stops if the last group is empty. The "natural" number of scores and groups therefore varies from distribution to distribution. However, experience shows that in the application to most citation distributions the use of three or four classes has satisfactory results. If more "fine-tuning" is required, the number of classes can, of course, be arbitrarily extended by adding further classes and the scale can thus be optionally refined. In order to obtain the four classes we proceed as follows.

In this empirical section we first determine the first three scores on basis of the ranked sample ($\beta_0 = 0$ by definition). Class 1 is formed by interval $[\beta_0, \beta_1)$. Its elements are less frequently cited than the average, those of categories 2 through 4 (intervals $[\beta_1, \beta_2)$, $[\beta_2, \beta_3)$ and $[\beta_3, \infty)$) are more cited than the average. Papers of the latter ones were called 'fairly cited', 'remarkably cited' and 'outstandingly cited', respectively, while the elements of the first class were called 'poorly cited' (cf., Glänzel & Schubert, 1988a, 1988b). In this manner, the original (theoretically infinite but practically finite) distribution is reduced to a finite distribution keeping essential properties of the original one and taking as many values as classes are used. Then we are ready to analyse the statistical properties of the empirical classes obtained from procedure described above.

In order to apply this method to empirical data all "citable" papers indexed in the 1980 volume of the *Science Citation Index* of Thomson-ISI (Philadelphia, PA, USA) have been collected. The set of "citable" papers includes the document type *article*, *letter*, *note* and *review*. The data set includes about 450,000 papers that have been assigned to subfields according to the Leuven/Budapest classification scheme consisting of 12 major fields and 60 subfields in the sciences (cf. Glänzel & Schubert, 2003). Citation counts have been determined on the basis of an item-by-item procedure using special identification-keys made up of bibliographic data elements such as part of the first author's name, the publication year, the volume number and first page of publication. Citations to each paper have been cumulated from the publication year till all individual years in the period 1980–2000. Finally citations have been aggregated at the level of the sixty subfields and for all fields combined.

Fig. 1 shows the evolution of relative group sizes on basis of cumulative citation rates received by the papers published in 1980 in all fields combined. The shares of the classes in the total is stabilising very soon after publication. The strikingly 'irregular' behaviour in the publication year 1980 could be explained with the fact that this year is still incomplete for most publications; a paper published in November or December has obviously less chance to be cited in the same year than a paper published in the first quarter of the year. This might also have effect on citation patterns in the subsequent year. However, already in 1983, that is, in the third year after publication, class shares do not change essentially any more. The class sizes form an extremely skewed distribution; 20 years after publication about one quarter (74.7%) of all papers received less-than-average cited, a bit less than one fifth (18.5%) are fairly cited,

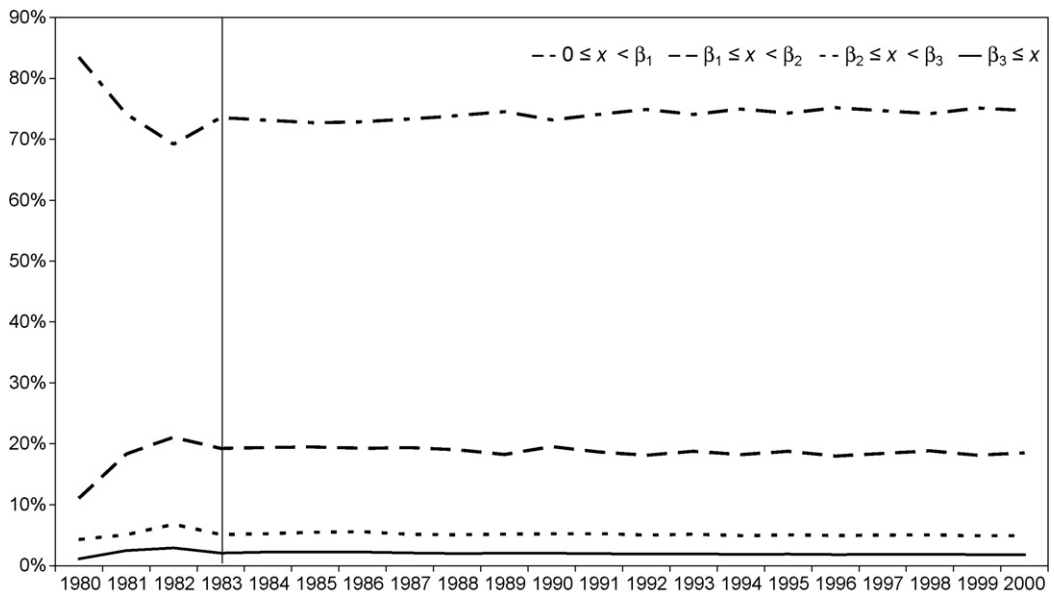


Fig. 1. Evolution of relative group sizes on basis of cumulative citation rates received by papers published in 1980 in the period 1980–2000 (all fields combined).

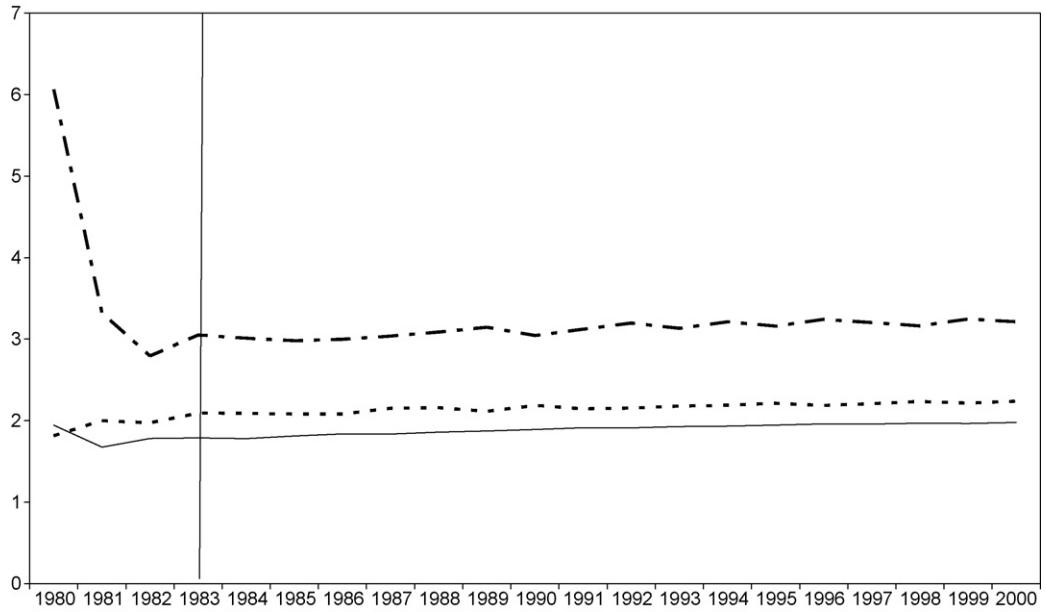


Fig. 2. Evolution of β_k/β_{k-1} ratios on basis of cumulative citation rates received by papers published in 1980 in the period 1980–2000 and in all fields combined ($k=2$ dashed line, $k=3$ dotted line, $k=4$ solid line).

practically one twentieth (4.9%) was remarkably cited and roughly 2% (1.8%) attracted outstanding citation impact. The class-size evolution in the 60 subfields parallels these patters. Stabilisation can be observed from the fourth year on and the sizes themselves are of similar order in most disciplines, too. The approximation 75% (size of class 1), 18% for class 2, 5% for class 3 and 2% for class 4 can be used as a rule of thumb for most disciplines, but should – as all informetric laws – be applied to research evaluation with the utmost caution. A more detailed discussion of the group properties of 12 out of the 60 subfields will form a part of the following section.

Fig. 2, which presents the evolution of β_k/β_{k-1} ratios for the papers published in 1980 in all fields combined, shows similar patterns of early stabilisation. For this exercise we have determined also the score β_4. The evolution of ratios shows a similar picture of stability beginning with the third year after publication. Although the ratios of subsequent score are similar in the individual subfields, the scores themselves vary considerably among the disciplines. Both the class sizes and the β_k/β_{k-1} ratios do not depend on parameter N (see previous section). Since citation processes converge quite slowly to their limiting distribution (cf. Glänzel & Schoepflin, 1995), the time stability of the two statistics shown in Figs. 1 and 2 indicates that N is the time-dependent whereas α does not depend on time.

4. Further results

On basis of the 1980 volume of the Science Index characteristic scores and scales were calculated for each individual citation window 1980 (1 year) through 1980–2000 (21 years). Twelve subfields out of the total of 60 subfields have been selected. Every subfield represents one major field each. The selection is presented in Table 2.

The sizes of these subfields range from $n=3727$ in *aquatic sciences* to $n=19{,}603$ in *pharmacology and toxicology*. All relevant statistics based on all possible citation windows have been determined; in particular the characteristic scores β_k (for $k=1,\ldots,4$) and the shares of the classes in the total $(\nu_{k-1}-\nu_k)/n$ for $k=1,\ldots,3$ and ν_3/n for $k=4$, where these shares are denoted by 'class_k'. Here we just mention in passing that $\Sigma_k\ \text{class}_k=\nu_0/n=1$. In addition, the parameter $q=(1-1/\alpha)^{\alpha}$ has been estimated according to the property (13), namely as $\tilde{q}=(\nu_k)^{1/k}$ for $k\geq 2$. The mean of the three $\tilde{q}$ values were used as the final estimator of the parameter q for the individual subfields. This estimator is denoted by $\bar{q}$. The corresponding parameter α of the assumed Pareto distribution can be found in Table 1. All above-mentioned statistics for the full citation period 1980–2000 are presented in Table 3.

The statistics presented in Table 2 substantiate that the Pareto approximation works sufficiently well for most subfields. While this approximation provides quite stable results, e.g., for the $\tilde{q}$ estimates of the subfields A4, M4 and

Table 2
List of selected subfields, subfield codes and majors fields to which the subfields are assigned

Code	Subfield	Major field
A4	Food and animal science and technology	Agriculture and environment
Z2	Aquatic sciences	Biology (organismic and supraorganismic level)
B1	Biochemistry/biophysics/molecular biology	Biosciences (general, cellular and subcellular biology; genetics)
R4	Pharmacology and toxicology	Biomedical research
I5	Immunology	Clinical and experimental medicine I (general and internal medicine)
M4	Ophthalmology/otolaryngology	Clinical and experimental medicine II (non-internal medicine specialties)
N1	Neurosciences and psychopharmacology	Neuroscience and behaviour
C1	Analytical, inorganic and nuclear chemistry	Chemistry
P6	Physics of solids, fluids and plasmas	Physics
G1	Astronomy and astrophysics	Geosciences and space sciences
E2	Electrical and electronic engineering	Engineering
H1	Applied mathematics	Mathematics

G1, the $\tilde{q}$ values of the other subfields seem to tendentiously increase or decrease with growing index k. These trends clearly show that the Pareto model can be considered only a rough approximation. The corresponding α parameter ranges between $\alpha = 2.0$ for B1 (biochemistry/biophysics/molecular biology) and 3.5 for M4 (ophthalmology/otolaryngology).

According to the theoretical considerations of the previous section, group sizes do not depend on the parameter N while the characteristic scores are a linear function of this parameter. The corresponding scores of samples with similar $\alpha(q)$ values might therefore considerably differ as the examples H1 (applied mathematics) and B1 (biochemistry/biophysics/molecular biology) show. For instance, a paper, which has received 50 citations in the period 1980–2000, would classify as 'outstandingly cited' in H1 (applied mathematics), but only 'remarkably cited' in C1 (analytical, inorganic and nuclear chemistry) and even just 'fairly cited' in I5 (immunology).

In principle, we can distinguish four basic types according to their citation standard (lower or higher threshold values β_k) and to their class-size distributions. A classification of these properties is presented in Table 4. While the first characteristics depend on both parameters, the latter distribution depends on $\alpha(q)$ alone. Eight of the selected subfield show specific profiles; the pairs M4 and A4, N1 and G1, H1and E2 as well as B1 and I5 have similar

Table 3
Characteristic scores and scales for 12 subfields based on the 21-year citation windows 1980–2000 (n = sample size, class$_k$ = kth class size/n, $\tilde{q} = (\nu_k)^{1/k}$, $\bar{q}$ = mean($\tilde{q}$))

k	A4($\bar{q} = 0.291, n = 8531$)			Z2($\bar{q} = 0.298, n = 3727$)			B1($\bar{q} = 0.257, n = 28,161$)			R4($\bar{q} = 0.284, n = 19,603$)		
	β_k	Class$_k$	$\tilde{q}$	β_k	Class$_k$	$\tilde{q}$	β_k	Class$_k$	$\tilde{q}$	β_k	Class$_k$	$\tilde{q}$
1	9.62	0.710	n/a	18.42	0.700	n/a	29.79	0.738	n/a	16.64	0.719	n/a
2	26.57	0.205	0.290	46.64	0.207	0.300	85.85	0.196	0.262	46.63	0.199	0.281
3	52.67	0.060	0.292	87.21	0.068	0.304	196.55	0.051	0.257	93.71	0.058	0.285
4	92.63	0.025	0.291	159.86	0.024	0.289	437.63	0.016	0.250	166.15	0.023	0.286
k	I5($\bar{q} = 0.262, n = 8811$)			M4($\bar{q} = 0.310, n = 5342$)			N1($\bar{q} = 0.294, n = 9691$)			C1($\bar{q} = 0.288, n = 16,508$)		
	β_k	Class$_k$	$\tilde{q}$	β_k	Class$_k$	$\tilde{q}$	β_k	Class$_k$	$\tilde{q}$	β_k	Class$_k$	$\tilde{q}$
1	27.00	0.741	n/a	11.66	0.692	n/a	27.46	0.713	n/a	14.24	0.711	n/a
2	79.17	0.191	0.259	30.05	0.211	0.308	74.15	0.202	0.287	37.30	0.206	0.289
3	174.06	0.050	0.262	54.52	0.067	0.311	142.38	0.058	0.292	72.60	0.060	0.288
4	335.88	0.019	0.266	85.76	0.030	0.311	233.90	0.027	0.301	129.39	0.023	0.285
k	P6($\bar{q} = 0.262, n = 14,330$)			G1($\bar{q} = 0.286, n = 5472$)			E2($\bar{q} = 0.254, n = 11,498$)			H1($\bar{q} = 0.254, n = 6675$)		
	β_k	Class$_k$	$\tilde{q}$	β_k	Class$_k$	$\tilde{q}$	β_k	Class$_k$	$\tilde{q}$	β_k	Class$_k$	$\tilde{q}$
1	14.10	0.743	n/a	23.24	0.712	n/a	7.36	0.759	n/a	6.51	0.755	n/a
2	41.36	0.187	0.257	63.23	0.207	0.288	25.10	0.176	0.241	21.65	0.181	0.245
3	88.34	0.051	0.264	127.96	0.058	0.285	55.85	0.047	0.255	49.66	0.047	0.255
4	167.20	0.019	0.266	232.95	0.023	0.285	104.87	0.019	0.265	98.24	0.018	0.263

Table 4
Four basis types of characteristic scores and scales according to thresholds and class sizes

Properties	Lower citation standard	Higher citation standard
Less skewed class-size distribution (α, q large)	M4, A4	N1, G1
More skewed class-size distribution (α, q small)	H1, E2	B1, I5

Table 5
Time dependence of parameter N as reflected by the change of characteristic scores in time

	Field							
	M4		A4		H1		E2	
	1983	2000	1983	2000	1983	2000	1983	2000
β_1	2.55	11.66	2.24	9.62	1.35	6.51	2.23	7.36
β_2	6.41	30.05	5.96	26.57	4.00	21.65	7.20	25.10
β_3	11.08	54.52	9.46	52.67	6.52	49.66	14.57	55.85
β_4	16.60	85.76	14.30	92.63	10.22	98.24	23.30	104.87
	Field							
	N1		G1		B1		I5	
	1983	2000	1983	2000	1983	2000	1983	2000
β_1	6.74	27.46	6.72	23.24	17.09	29.79	8.95	27.00
β_2	16.11	74.15	16.12	63.23	46.14	85.85	22.32	79.17
β_3	29.04	142.38	28.78	127.96	96.57	196.55	44.18	174.06
β_4	45.52	233.90	43.83	232.95	182.98	437.63	78.19	335.88

profiles each where the distribution characteristics of the first and fourth and the second and third, respectively, can be considered opposite to each other. Most subfields are, however, somewhere in between these 'extreme' profiles. The class sizes of the individual subfields somewhat differ from the above-mentioned rule of thumb (75–18–5–2%), although the deviations from this 'standard' are not dramatic. However, the variation among the characteristic scores of the individual subfields is – according to the expectation on basis of subject specific citation standards – quite considerable.

Finally, we have a look at the time dependence of parameter N. The characteristic scores depend on both parameters N and α (see Eq. (11)). In Table 5, the characteristic scores of the eight subfields shown in Table 4 are presented for the two periods 1980–1983 (4-year window) and 1980–2000 (21-year window). The first four disciplines represent a lower citation standard (see upper part of Table 5) whereas the second group in the lower part of Table 5 stands for higher standards. The different 'growth rates' of the β_k values can be explained with the different ageing of the subfields (cf. Glänzel & Schoepflin, 1995). The deviation of the growth in the mathematics discipline from that in the engineering subfield is most striking (see Table 5).

5. Conclusions

The characteristic scores and scales defined by Glänzel & Schubert (1988a, 1988b) proved an interesting self-adjusting informetric tool that can be used for evaluative scientometric purposes as well. It can, for instance, be applied to journal and subject analysis. The identification of publication subsets according to their citation impact and gauging the mean citation rates of given subsets against a larger characteristic scale are likewise possible. Both scores and class sizes have interesting mathematical properties if a Pareto distribution for the underlying citation distribution is assumed. The size of the classes defined by the characteristic scores as well as the ratios of subsequent scores proved stable beyond an initial citation period of about 3 years. This observation provides on one hand empirical evidence for the time independence of the characteristic parameter (α) of the Pareto distribution and, on the other hand, it shows that the particular choice of the citation windows is – except for a short initial period – not important for class sizes. This does, of course, not imply that the elements of these classes, that is, the individual papers also form stable clusters.

The statement "once highly (poorly) cited always highly (poorly) cited" does not hold. While the size of the classes does not essentially change as time elapses, the composition of the classes might do.

The advantage of being able to determine these characteristic classes on basis of a shorter period and than to apply them to a larger citation window is maybe the most interesting property. The second important property is the 75–18–5–2 property, namely, that about 75% of the papers published in 1980 were poorly cited, 5% and 2% of the papers were 'remarkably cited' and 'outstandingly cited', respectively. However, this reflects the situation in 1980. Scientific communication has considerable changed during the last two decades; research collaboration and consequently also co-authorship has intensified, individual publication activity and citation impact have generally increased (see Persson, Glänzel, & Danell, 2004). The fact that the journal and disciplinary citation impact is growing but not all subjects are concerned to the same extent, points also to structural changes. The following three subfields may serve just as an example. The mean citation rate of subfield P6 (physics of solids, fluids and plasmas) based on a 3-year citation windows has increased from the publication year 1980 to the publication year 2002 by 168.4%. The increase of the 3-year citation impact in R4 (pharmacology and toxicology) was by far more moderate; it amounts to 38.2% in 2002. By contrast, the change of 6.1% in E2 (electrical and electronic engineering) is almost negligible. One can therefore expect that the above-mentioned 75–18–5–2 rule could somewhat change if the exercise is repeated, say, for papers published in 2000.

Besides its informetric value the method of characteristic scores and scales has implications also for research evaluation. The most important one is that the method provides a rule where thresholds could be set for the identification of highly cited papers within a scientific discipline and that this can be done whether on basis of characteristic scores which increase with growing citation window or on basis of the class sizes which do not change essentially beyond a certain initial citation period. Unknown scores can even be estimated from the lowest citation rate in the set of 2% and 7% (=2 + 5%) most cited papers, respectively. This method can easily be used as reference standard as well. Thus, the high-end of the national or institutional citation distribution can be gauged against the field standard. The share of subject relevant papers of a given country, region or institution found in one of the 'upper classes' can thus enlighten the user on the possible correspondence of the high-end of their citation distribution with the reference standard.

The Pareto approach used in the theoretic section provides powerful tools for evaluative bibliometrics; the α parameter (and its derivative q) characterises both the class-size distributions and the slope of the underlying citation distribution. Jointly with an 'impact measure' it can be used to distinguish four basic types of citation standard. Furthermore, the characteristic parameter proved to be time-independent, that is, it can be applied to any larger citation window once estimated for a given initial period.

It should be stressed again that the results obtained from this methods are of statistical value. The method of characteristic scores and scales should be applied to publication sets of reasonable size, that is, if the grouping procedure is complete for the ranked observations. It is certainly not designed for the assessment of individual papers, either, since those do not necessarily remain in their classes as time elapses, and might, therefore, be replaced by others.

Acknowledgements

The author wishes to thank Professor Jacqueline Leta for her help in applying the Leuven-Budapest subject classification scheme to the 1980 volume of the Science Citation Index. He also wishes to thank the anonymous referees for their corrections, critical comments and valuable suggestions.

References

Abt, H. A. (1981). Long-term citation histories of astronomical papers. *Publications of the Astronomical Society of the Pacific*, *93*(552), 207–210.

Burrell, Q. L. (1990). Using the gamma-Poisson model to predict library circulations. *Journal of the American Society for Information Science*, *41*(3), 164–170.

Burrell, Q. L. (2001). Stochastic modelling of the first-citation distribution. *Scientometrics*, *52*(1), 3–12.

Burrell, Q. L., & Cane, V. R. (1982). The analysis of library data. *Journal of the Royal Statistical Society Series A*, *145*, 439–471.

Egghe, L., & Rao, I. K. R. (2001). Theory of first-citation distributions and applications. *Mathematical and Computer Modelling*, *34*(1–2), 81–90.

Garfield, E. (1980). Premature discovery or delayed recognition. Why? *Current Contents*, *21*, 5–10.

Garfield, E. (1996). How can impact factors be improved? *British Medical Journal*, *313*(7054), 411–413.

Garfield, E. (1998a). Long-term vs. short-term journal impact: Does it matter? *Scientist*, *12*(3), 11–12.

Garfield, E. (1998b). Long-term vs. short-term impact. Part II. *Scientist*, *12*(14), 12–13.

Glänzel, W. (1992). On some stopping times on citation processes, from theory to indicators. *Information Processing and Management*, *28*(1), 53–60.

Glänzel, W. (2005). Methodological questions of subject characteristics in long-term citation analysis. An informetric study with implication for research evaluation. In *Paper presented at the 10th Nordic workshop on bibliometrics, informetrics and research policy held in Stockholm*.

Glänzel, W., & Garfield, E. (2004). The myth of delayed recognition. *The Scientist*, *18*(11), 8–9.

Glänzel, W., & Moed, H. F. (2002). Journal impact measures in bibliometric research. *Scientometrics*, *53*(2), 171–193.

Glänzel, W., & Schoepflin, U. (1994). A stochastic model for the ageing of scientific literature. *Scientometrics*, *30*(1), 49–64.

Glänzel, W., & Schoepflin, U. (1995). A bibliometric study on aging and reception processes of scientific literature. *Journal of Information Science*, *21*(1), 37–53.

Glänzel, W., & Schubert, A. (1988a). Characteristic scores and scales in assessing citation impact. *Journal of Information Science*, *14*(2), 123–127.

Glänzel, W., & Schubert, A. (1988b). Theoretical and empirical studies of the tail of scientometric distributions. In L. Egghe, & R. Rousseau (Eds.), *Informetrics Vols. 87/88*, (pp. 75–83). Elsevier Science Publisher B.V.

Glänzel, W., & Schubert, A. (1995). Predictive aspects of a stochastic model for citation processes. *Information Processing and Management*, *31*(1), 69–80.

Glänzel, W., & Schubert, A. (2003). A new classification scheme of science fields and subfields designed for scientometric evaluation purposes. *Scientometrics*, *56*(3), 357–367.

Glänzel, W., Telcs, A., & Schubert, A. (1984). Characterization by truncated moments and its application to Pearson-type distributions. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, *66*, 173–183.

Glänzel, W., Schlemmer, B., & Thijs, B. (2003). Better late than never? On the chance to become highly cited only beyond the standard bibliometric time horizon. *Scientometrics*, *58*(3), 571–586.

Gumbel, E. J. (1958). *Statistics of extremes*. New York: Columbia University Press.

Irwin, J. O. (1975a). Generalized Waring distribution. Part 1. *Journal of the Royal Statistical Society Series A*, *138*, 18–31.

Irwin, J. O. (1975b). Generalized Waring distribution. Part 2. *Journal of the Royal Statistical Society Series A*, *138*, 204–227.

Irwin, J. O. (1975c). Generalized Waring distribution. Part 3. *Journal of the Royal Statistical Society Series A*, *138*, 374–384.

Johnson, N. L., Kotz, S., & Balakrishnan, N. (1994). (2nd ed.). *Continuous univariate distributions Vol. 1*, New York: John Wiley & Sons, Inc.

Karlis, D., & Xekalaki, E. (2005). Mixed Poisson distributions. *International Statistical Review*, *73*(1), 35–58.

Persson, O., Glänzel, W., & Danell, R. (2004). Inflationary bibliometric values: The role of scientific collaboration and the need for relative indicators in evaluative studies. *Scientometrics*, *60*(3), 421–432.

Rousseau, R. (1994). Double exponential models for 1st-citation processes. *Scientometrics*, *30*(1), 213–227.

Schubert, A., & Glänzel, W. (1983). Statistical reliability of comparisons based on the citation impact of scientific publications. *Scientometrics*, *5*(1), 59–74.

van Raan, A. F. J. (2004). Sleeping beauties in science. *Scientometrics*, *59*(3), 467–472.

Xekalaki, E. (1983). The univariate generalized Waring distribution in relation to accident theory—proneness, spells or contagion. *Biometrics*, *39*(4), 887–895.

Available online at www.sciencedirect.com
ScienceDirect
Journal of Informetrics 1 (2007) 8–15
Journal of INFORMETRICS An International Journal
www.elsevier.com/locate/joi

Finding scientific gems with Google's PageRank algorithm

P. Chen[a,*], H. Xie[b,c], S. Maslov[c], S. Redner[a]

[a] *Center for Polymer Studies and Department of Physics, Boston University, Boston, MA 02215, United States*
[b] *New Media Lab, The Graduate Center, CUNY, New York, NY 10016, United States*
[c] *Department of Condensed Matter Physics and Materials Science, Brookhaven National Laboratory, Upton, NY 11973, United States*

Received 27 April 2006; received in revised form 12 June 2006; accepted 12 June 2006

Abstract

We apply the Google PageRank algorithm to assess the relative importance of all publications in the Physical Review family of journals from 1893 to 2003. While the Google number and the number of citations for each publication are positively correlated, outliers from this linear relation identify some exceptional papers or "gems" that are universally familiar to physicists.

PACS: 02.50.Ey; 05.40.-a; 05.50.+q; 89.65.-s

Keywords: Google PageRank algorithm; Scientific gems; Physical Review; Citations

1. Introduction

With the availability of electronically available citation data, it is now possible to undertake comprehensive studies of citations that were unimaginable just a few years ago. In most previous studies of citation statistics, the metric used to quantify the importance of a paper is its number of citations. In terms of the underlying citation network, in which nodes represent publications and directed links represent a citation from a *citing* article to a *cited* article, the number of citations to an article translates to the in-degree of the corresponding node. The distribution of in-degree for various citation data sets has a broad tail (Price, 1965, 1976) that is reasonably approximated by a power law (Laherrère & Sornette, 1998; Redner, 1998, 2005).

While the number of citations is a natural measure of the impact of a publication, we probably all have encountered examples where citations do not seem to provide a full picture of the influence of a publication. We are thus motivated to study alternative metrics that might yield a truer measure of importance than citations alone. Such a metric already exists and is provided by the Google PageRank (Brin & Page, 1998) or similar algorithms proposed for the analysis of social (Bonacich, 1972, 1987) and information (Kleinberg, 1999) networks. By simulating random traffic on a network these algorithms calculate the importance of papers in a self-consistent fashion. They naturally take into account the following two factors: (1) the effect of receiving a citation from a more important paper should be greater that that coming from a less popular one; (2) citation links coming from a paper with a long reference list should count less than those coming from one with a short list. In other words, the importance of a paper should be divided over a number of references that inspired this line of research.

* Corresponding author.
E-mail addresses: patrick@bu.edu (P. Chen), hxie@bnl.gov (H. Xie), maslov@bnl.gov (S. Maslov), redner@bu.edu (S. Redner).

1751-1577/$ – see front matter © 2006 Published by Elsevier Ltd.
doi:10.1016/j.joi.2006.06.001

A variant of the PageRank algorithm was recently applied to better calibrate the impact factor of scientific journals (Bollen, Rodriguez, & Van de Sompel, 2006). In this work, we apply Google PageRank to the Physical Review citation network with the goal of measuring the importance of individual scientific publications published in the APS journals. This network consists of 353,268 nodes that represent all articles published in the Physical Review family of journals from the start of publication in 1893 until June 2003, and 3,110,839 links that represent all citations *to* Physical Review articles *from* other Physical Review articles. As found previously (Redner, 2005), these internal citations represent 1/5 to 1/3 of all citations for highly-cited papers. This range provides a sense of the degree of completeness of the Physical Review citation network.

With the Google PageRank approach, we find a number of papers with a modest number of citations that stand out as exceptional according to the Google PageRank analysis. These exceptional publications, or gems, are familiar to almost all physicists because of the very influential contents of these articles. Thus, the Google PageRank algorithm seems to provide a new and useful measure of scientific quality.

2. The Google PageRank algorithm

To set the stage for our use of Google PageRank to find scientific gems, let us review the elements of the PageRank algorithm. Given a network of N nodes $i = 1, 2, \ldots, N$, with directed links that represent references from an initial (citing) node to a target (cited) node, the Google number G_i for the ith node is defined by the recursion formula (Brin & Page, 1998):

$$G_i = (1-d)\sum_{j \text{ nn } i} \frac{G_j}{k_j} + \frac{d}{N}. \tag{1}$$

Here the sum is over the neighboring nodes j in which a link points to node i. The first term describes propagation of the probability distribution of a random walk in which a walk at node j propagates to node i with probability $1/k_j$, where k_j is the out-degree of node j. The second term describes the uniform injection of probability into the network in which each node receives a contribution d/N at each step.

Here d is a free parameter that controls the performance of the Google PageRank algorithm. The prefactor $(1-d)$ in the first term gives the fraction of random walks that continue to propagate along the links; a complementary fraction d is uniformly re-injected into the network, as embodied by the second term.

We suggest that the Google number G_i of paper i, defined by Eq. (1), is a better measure of importance than the number of citations alone in two aspects: (i) being cited by influential papers contributes more to the Google number than being cited by unimportant papers; (ii) being cited by a paper that itself has few references gives a larger contribution to the Google number than being cited by a paper with hundreds of references. The Google number of a paper can be viewed as a measure of its influence that is then equally exported to all of its references. The parameter $d > 0$ prevents all of the influence from concentrating on the oldest papers.

In the original Google PageRank algorithm of Brin and Page (1988), the parameter d was chosen to be 0.15. This value was prompted by the anecdotal observation that an individual surfing the web will typically follow of the order of 6 hyperlinks, corresponding to a leakage probability $d = 1/6 \simeq 0.15$, before becoming either bored or frustrated with this line of search and beginning a new search. In the context of citations, we conjecture that entries in the reference list of a typical paper are collected following somewhat shorter paths of average length 2, making the choice $d = 0.5$ more appropriate for a similar algorithm applied to the citation network. The empirical observation justifying this choice is that approximately 50% of the articles[1] in the reference list of a given paper A have at least one citation $B \to C$ to another article C that is also in the reference list of A. Assuming that such "feed-forward" loops result from authors of paper A following references of paper B, we estimate the probability $1-d$ to follow this indirect citation path to be close to 0.5.

To implement the Google PageRank algorithm for the citation network, we start with a uniform probability density equal to $1/N$ at each node of the network and then iterate Eq. (1). Eventually a steady state set of Google numbers for each node of the network is reached. These represent the occupation probabilities at each node for the random-walk-like process defined by Eq. (1). Finally, we sort the nodal Google numbers to determine the Google rank of each node. It

[1] The actual fraction of "followed citations" (such as B in an A → B, A → C, and B → C loop) is 42% for the entire dataset and 51% for papers published during the last 4 years.

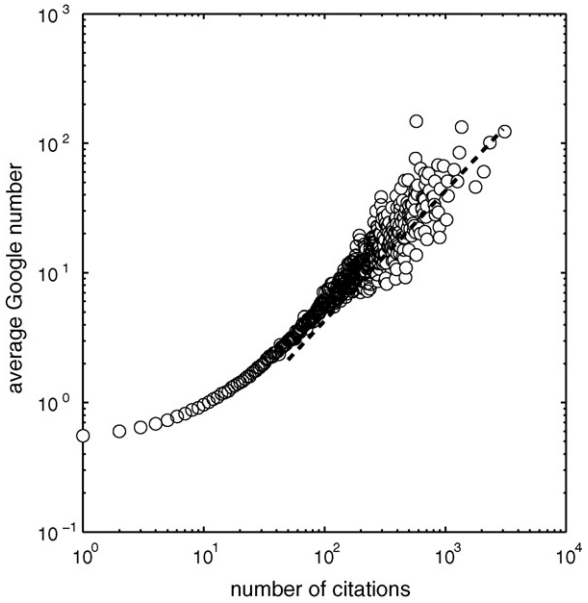


Fig. 1. Average Google number $\langle G(k)\rangle$ vs. number of citations k. The dashed line of slope 1 is a guide for the eye.

is both informative and entertaining to compare the Google rank with the citation (in-degree) rank of typical and the most important publications in Physical Review.

3. Google's PageRank for Physical Review

Fig. 1 shows the average Google number $\langle G(k)\rangle$ for publications with k citations as a function of k. For small k, there are many publications with the same number of citations and the dispersion in $G(k)$ is small. Correspondingly, the plot of $\langle G(k)\rangle$ versus k is smooth and increases linearly with k for $k \gtrsim 50$. Thus, the average Google number and the number of citations represent similar measures of popularity, a result that has been observed previously (Fortunato, Boguna, Flammini, & Menczer; Fortunato, Flammini, & Menczer). In fact, the citation and Google number distributions are

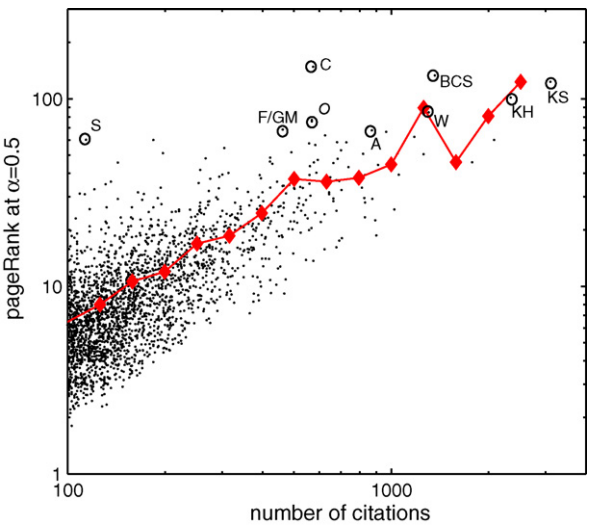


Fig. 2. Individual outlier publications. The scatter plot of the Google number vs. the number of citations. The top-10 Google-ranked papers are identified by author(s) initials (see Table 1). As a guide to the eye, the solid curve is logarithmically binned average of the data of $\langle G(k)\rangle$ vs. k in Fig. 1.

Table 1
The top-10 Google-ranked publications when $d = 0.5$

Google rank	Google # ($\times 10^{-4}$)	Cite rank	# cites	Publication				Title	Author(s)
1	4.65	54	574	PRL	10	531	1963	Unitary symmetry and leptonic...	N. Cabibbo
2	4.29	5	1364	PR	108	1175	1957	Theory of superconductivity	J. Bardeen, L. Cooper, and J. Schrieffer
3	3.81	1	3227	PR	140	A1133	1965	Self-consistent equations...	W. Kohn and L.J. Sham
4	3.17	2	2460	PR	136	B864	1964	Inhomogeneous electron gas	P. Hohenberg and W. Kohn
5	2.65	6	1306	PRL	19	1264	1967	A model of leptons	S. Weinberg
6	2.48	55	568	PR	65	117	1944	Crystal statistics I	L. Onsager
7	2.43	56	568	RMP	15	1	1943	Stochastic problems in...	S. Chandrasekhar
8	2.23	95	462	PR	109	193	1958	Theory of the Fermi interaction	R.P. Feynman and M. Gell-Mann
9	2.15	17	871	PR	109	1492	1958	Absence of diffusion in...	P.W. Anderson
10	2.13	1853	114	PR	34	1293	1929	The theory of complex spectra	J.C. Slater

qualitatively similar, further indicating that citations and Google numbers are, on the average, similar measures of importance.

However, for large k, much more interesting behavior occurs. When k is sufficiently large, there is typically only one publication with k citations. Of particular interest are the extreme outliers with respect to the linear behavior of Fig. 1. The 10 articles with the highest Google numbers are marked with large circles in Fig. 2 on the background of Google number versus the number of citations for all publications (dots). The top-10 papers are identified by author initials (see Table 1). Table 1 also lists the number of citations and the citation rank of these publications. While several of the highest-cited Physical Review papers appear on this list, there are also several more modestly-cited papers that are highly ranked according to the Google algorithm.

The disparity between the Google rank and citation rank arises because, as mentioned in the previous section, the former involves both the in-degree as well as the Google PageRank of the neighboring nodes. According to the Google algorithm of Eq. (1), a citing publication ("child") j contributes a factor $\langle G_j/k_j\rangle$ to the Google number of its parent paper i. Thus, for a paper to have a large Google number, its children should be important (large G_j), and also each child should have a small number of parents (small out-degree k_j). The latter ensures that the Google contribution of a child is not strongly diluted.

With this perspective, let us compare the statistical measures of the two articles "Unitary Symmetry and Leptonic Decays", Phys. Rev. Lett. 10, 531 (1963) by N. Cabibbo (C) and "Self-Consistent Equations Including Exchange and Correlation Effects", Phys. Rev. 140, A1133 (1965) by W. Kohn & L. J. Sham (KS). The former has the highest Google number of all Physical Review publications, while the latter is the most cited. The high Google rank of C stems from the fact that that value of $\langle G_j/k_j\rangle = 1.52 \times 10^{-6}$ for the children of C is an order of magnitude larger than the corresponding value $\langle G_j/k_j\rangle = 2.31 \times 10^{-7}$ for the children of KS. This difference more than compensates for the factor 5.6 difference in the number of citations to these two articles (3227 for KS and 574 for C as of June 2003). Looking a little deeper, the difference in $\langle G_j/k_j\rangle$ for C and KS stems from the denominator; the children of C have 15.6 citations an average, while the children of KS are slightly "better" and have 18.4 citations on average. However, the typical child of C has fewer references than a child of KS and a correspondingly larger contribution to the Google number of C.

The remaining research articles on the top-10 Google-rank list but outside the top-10 citation list are easily recognizable as seminal publications. For example, Onsager's 1944 paper presents the exact solution of the two-dimensional Ising model; both a calculational *tour de force*, as well as a central development in the theory of critical phenomena. The paper by Feynman and Gell-Mann introduced the *V-A* theory of weak interactions that incorporated parity non-conservation and became the "standard model" of weak interactions. Anderson's paper, "Absence of Diffusion in Certain Random Lattices" gave birth to the field of localization and is cited by the Nobel prize committee for the 1977 Nobel prize in physics.

The last entry in the top-10 Google-rank list, "The Theory of Complex Spectra", by J. C. Slater (S) is particularly striking. This article has relatively few citations (114 as of June 2003) and a relatively low citation rank (1853th),

Table 2
The remaining top-100 Google-ranked papers when $d = 0.5$ in which the ratio of Google rank to citation rank is greater than 10

Google rank	Google # ($\times 10^{-4}$)	Cite rank	# cites	Publication				Title	Author(s)
1	4.65	54	574	PRL	10	531	1963	Unitary symmetry and leptonic...	N. Cabibbo
8	2.23	95	462	PR	109	193	1958	Theory of the fermi interaction	R.P. Feynman and M. Gell-Mann
10	2.13	1853	114	PR	34	1293	1929	The theory of complex spectra	J.C. Slater
12	2.11	712	186	PR0	43	804	1933	On the constitution of...	E. Wigner and F. Seitz
20	1.80	228	308	PR0	106	364	1957	Correlation energy of an ...	M. Gell-Mann and K. Brueckner
21	1.69	616	198	PRL	58	408	1987	Bulk superconductivity at ...	R.J. Cava, et al.
25	1.58	311	271	PRL	58	405	1987	Evidence for superconductivity ...	C.W. Chu, et al.
30	1.51	1193	144	PRL	10	84	1963	Photon correlations	R.J. Glauber
35	1.42	12897	39	PR0	35	509	1930	Cohesion in monovalent metals	J.C. Slater
49	1.21	1342	136	PR0	60	252	1941	Statistics of the two- ...	H.A. Kramers and G.H. Wannier
58	1.17	1433	135	PR0	81	440	1951	Interaction between the ...	C. Zener
59	1.17	5196	66	PR0	45	794	1934	Electronic energy bands in ...	J.C. Slater
60	1.16	2927	108	PRB	28	4227	1983	Electronic structure of ...	L.F. Mattheiss & D. R. Hamann
64	1.12	642	199	PR0	52	191	1937	The structure of electronic ...	G.H. Wannier
70	1.08	1653	130	PRL	10	518	1963	Classification of two-electron ...	J. Cooper, U. Fano, and F. Prats
72	1.06	1901	118	PR0	46	509	1934	On the constitution of ...	E. Wigner and F. Seitz
73	1.05	876	180	PR0	75	486	1949	The radiation theories of ...	F.J. Dyson
78	1.03	1995	119	PR0	109	1860	1958	Chirality invariance and ...	E. Sudarshan and R. Marshak
85	1.00	201853	3	PRB	22	5797	1980	Cluster formation in ...	H. Rosenstock and C. Marquardt
87	0.99	10168	48	PRL	6	106	1961	Population inversion and ...	A. Javan, W. Bennett, and D. Herriott
90	0.98	3231	86	PR0	79	350	1950	Antiferromagnetism ...	P.W. Anderson
92	0.97	1199	149	PR0	76	749	1949	The theory of positrons	R.P. Feynman

but its Google number 2.13×10^{-4} is only a factor 2.2 smaller than that of Cabibbo's paper! What accounts for this high Google rank? From the scientific standpoint, Slater's paper introduced the determinant form for the many-body wavefunction. This form is so ubiquitous in current literature that very few articles actually cite the original work when the Slater determinant is used. The Google PageRank algorithm identifies this hidden gem primarily because the average Google contribution of the children of S is $\langle G_j/k_j \rangle = 3.51 \times 10^{-6}$, which is a factor 2.3 larger than the contribution of the children of C. That is, the children of Slater's paper were both influential and Slater loomed as a very important father figure to his children.

The striking ability of the Google PageRank algorithm to identify influential papers can be seen when we consider the top-100 Google-ranked papers. Table 2 shows the subset of publications on the top-100 Google rank in which the ratio of Google rank to citation rank is greater than 10; that is, publications with anomalously high Google rank compared to their citation rank. This list contains many easily-recognizable papers for the average physicist. For example, the publication by Wigner and Seitz, "On the Constitution of Metallic Sodium" introduced Wigner-Seitz cells, a construction that appears in any solid-state physics text. The paper by Gell-Mann and Brueckner, "Correlation Energy of an Electron Gas at High Density" is a seminal publication in many-body theory. The publication by Glauber, "Photon Correlations", was recognized for the 2005 Nobel prize in physics. The Kramers-Wannier article, "Statistics of the Two-Dimensional Ferromagnet. Part I", showed that a phase transition occurs in two dimensions, contradicting the common belief at the time. The article by Dyson, "The Radiation Theories of Tomonaga, Schwinger, and Feynman", unified the leading formalisms for quantum electrodynamics and it is plausible that this publication would have earned Dyson the Nobel prize if it could have been shared among four individuals. One can offer similar rationalizations for the remaining articles in this table.

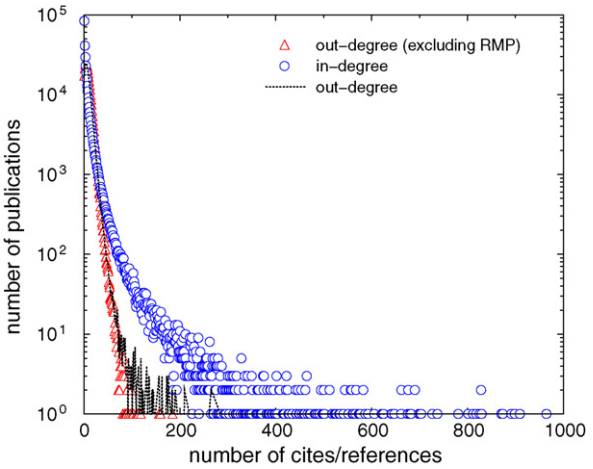


Fig. 3. The in-degree distribution (citations to) and out-degree distribution (the length of the reference list) for all Physical Review publications. The out-degree distribution is shown with and without the contribution of Reviews of Modern Physics. Eleven papers with more than 1000 citations are beyond the upper limit of the x-axis.

On the other hand, an apparent mistake is the paper by Rosenstock and Marquardt, "Cluster formation in two-dimensional random walks: Application to photolysis of silver halides" (RM). Notice that this article has only three citations! Why does RM appear among the top-100 Google-ranked publications? In RM, a model that is essentially diffusion-limited aggregation is introduced. Although these authors had stumbled upon a now-famous model, they focused on the kinetics of the system and apparently did not appreciate its wonderful geometrical features. This discovery was left to one of the children of RM—the famous paper by T. Witten and L. Sander, "Diffusion-Limited Aggregation, a Kinetic Critical Phenomenon" Phys. Rev. Lett. **47**, 1400 (1981), with 680 citations as of June 2003. Furthermore, the Witten and Sander article has only 10 references; thus a substantial fraction of its fame is exported to RM by the Google PageRank algorithm. The appearance of RM on the list of top-100 Google-ranked papers occurs precisely because of the mechanics of the Google PageRank algorithm in which being one of the few references of a famous paper makes a huge contribution to the Google number.

A natural question to ask is whether the Google rankings are robust with respect to the value of the free parameter d in the Google algorithm. As mentioned above, we believe that our *ad hoc* choice of $d = 0.5$ accounts in a reasonable way for the manner in which citations are actually made. For $d = 0.15$, as in the original Google algorithm, the Google rankings of highly-cited papers locally reorder to a considerable extent compared to the rankings for the case $d = 0.5$, but there is little global reordering. For example, all of the top-10 Google-ranked papers calculated with $d = 0.5$ remained among the top-50 Google-ranked papers for $d = 0.15$. Thus, up to this degree of variation, Google rankings are a robust measure. To further quantify it we measured the Spearman rank correlation of PageRank(d) with our choice of PageRank(0.5). Overall the correlation is very high. It varies between 0.98 and 1 for $0.1 < d < 0.9$. The rank correlation 0.91 with the number of citations is also quite high.

One can show that for $d \to 1$ Google rank almost exactly reduces to the citation rank. Indeed, in the extreme case of $d = 1$, the Google number of each node equals $1/N$. For $d \to 1$, we therefore write $d = 1 - \epsilon$, with $\epsilon \ll 1$, and also assume that there is a correspondingly small deviation of the Google numbers from $1/N$. Thus we write $G_i = \frac{1}{N} + \mathcal{O}(\epsilon)$. Substituting these into Eq. (1), we obtain

$$G_i = \epsilon \sum_j \frac{G_j}{k_j} + \frac{1-\epsilon}{N} \approx \frac{1}{N}\left[1 + \epsilon\left(\sum_j \frac{1}{k_j} - 1\right)\right] \tag{2}$$

To estimate the sum in Eq. (2), we use the fact that the out-degree distribution is relatively narrow (Fig. 3), especially if we exclude the broad tail that is caused by the contributions of review articles that appear in the Reviews of Modern Physics. While the mean in-degree and out-degrees are both close to 9 (and should be exactly equal for the complete citation network), the dispersion for the in degree is 23.15, while the dispersion for the out degree (excluding Reviews of Modern Physics) is 8.64.

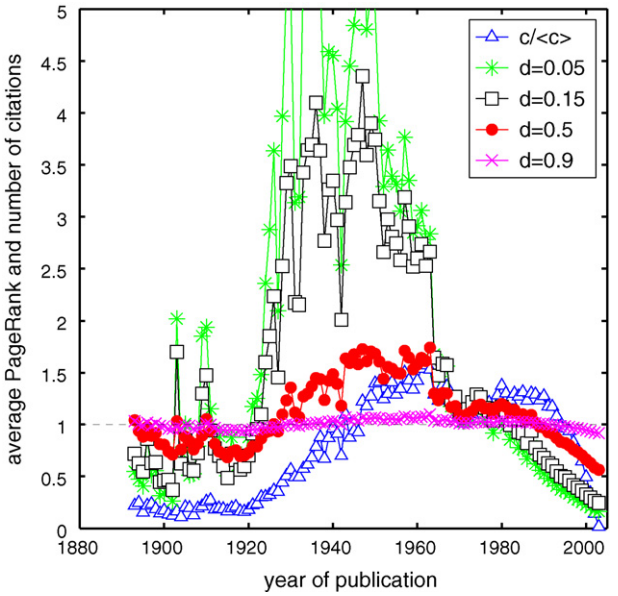


Fig. 4. Average value of PageRank vs. year of publication. Different symbols correspond to different values of d: 0.05—green stars, 0.15—black open squares, 0.5—red filled circles, 0.9—purple x's. Open blue triangles show the number citations c_i which, for proper comparison with the PageRank, is normalized by its average value ($\langle c \rangle = 8.08$). Dashed line at 1 gives the global average value of PageRank.

As a result of the sharpness of the out-degree distribution, the sum $\sum_{j \,\mathrm{nn}\, i} \frac{1}{k_j}$ for nodes with high in-degree is approximately equal to the in-degree c_i of node i times $\langle \frac{1}{k} \rangle$. With this assumption, Eq. (2) becomes

$$G_i = \frac{1}{N}\left[1 + \epsilon\left(c_i \left\langle \frac{1}{k} \right\rangle - 1\right)\right] . \tag{3}$$

That is, the leading correction to the limiting $d = 1$ result that $G_i = \frac{1}{N}$ is proportional to the in-degree (the number of received citations) c_i of each node. Thus, as $d \to 1$, the Google rank of each node is identical to its citation rank under the approximation that we neglect the effect of the dispersion of the out-degree in the citation network.

One final aspect to consider is whether PageRank gives an "unfair" advantage to older papers. Indeed long random walks on time-directed networks inevitably drift towards older papers (Xie et al., in preparation). Since the average length of the walk in the PageRank algorithm is given by $1/d$, this effect is especially pronounced for small values of d. To investigate this question, we plot in Fig. 4 the average value of normalized PageRank as a function of the year of publication for different values of d. For comparison we also include the same plot for the normalized number of citations. As one can see from this figure, for $d = 0.5$ the age variation is not great and is comparable to that in the number of citations. However, in agreement with our qualitative understanding, the relative weight assigned to older papers (those published in 1920–1960) increases for smaller values of d. Interestingly, the oldest papers in our dataset (those published before ∼1920 overall have below average PageRank as well as the number of citations.

4. Conclusions

We believe that protocols based on the Google PageRank algorithm hold a great promise for quantifying the impact of scientific publications. They provide a meaningful extension to traditionally-used importance measures, such as the number of citation of individual articles and the impact factor for journals as a whole. The PageRank algorithm implements, in an extremely simple way, the reasonable notion that citations from more important publications should contribute more to the rank of the cited paper than those from less important ones. Other ways of attributing a quality for a citation would require much more detailed contextual information about the citation itself.

The situation in citation networks is not that dissimilar from that in the World Wide Web, where hyperlinks contained in popular websites and pointing to your webpage would bring more Internet traffic to you and thus would contribute

substantially to the popularity of your own webpage. Scientists commonly discover relevant publications by simply following chains of citation links from other papers. Thus, it is reasonable to assume that the popularity or "citability" of papers may be well approximated by the random surfer model that underlies the PageRank algorithm. One meaningful difference between the WWW and citation networks is that citation links cannot be updated after publication, while WWW hyperlinks keep evolving together with the webpage containing them. Thus, scientific papers and their citations tend to age much more rapidly than active webpages. These differences could be taking into account by explicitly incorporating the effects of aging into the Page Rank algorithm (Xie et al., in preparation).

Acknowledgments

Two of us, (PC and SR) gratefully acknowledge financial support from the US National Science Foundation grant DMR0535503. SR also thanks A. Castro Neto, A. Cohen, and K. Lane for literature advice. Work at Brookhaven National Laboratory was carried out under Contract No. DE-AC02-98CH10886, Division of Material Science, U.S. Department of Energy.

References

Bollen, J., Rodriguez, M. A., & Van de Sompel, H. (2006). Journal status. *Scientometrics*, *69*(3), 669–687. This work was later put in the context of other relevant publications in: Ball, P. (2006). Prestige is factored into journal ratings. *Nature, 439*(7078), 770–771.

Bonacich, P. F. (1972). Factoring and weighting approaches to status scores and clique identification. *Journal of Mathematical Sociology*, *2*, 113–120.

Bonacich, P. F. (1987). Power and centrality: A family of measures. *American Journal of Sociology*, *92*, 1170–1182.

Brin, S., & Page, L. (1988). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, *30*, 107–117.

Fortunato, S., Boguna, M., Flammini, A., & Menczer, F. How to make the top ten: Approximating PageRank from in-degree. Cs.IR/0511016.

Fortunato, S., Flammini, A., & Menczer, F. (2006). Scale-free network growth by ranking. *Physics Review Letters*, *96*, 218701.

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM*, *46*, 604–632.

Laherrère, J., & Sornette, D. (1998). Stretched exponential distributions in nature and economy: "Fat tails" with characteristic scales. *European Physical Journal B*, *2*, 525–539.

Price, D. J. de Solla. (1965). Networks of Scientific Papers. *Science*, *149*, 510–515.

Price, D. J. de Solla. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, *27*, 292–306.

Redner, S. (1998). How popular is your paper? An empirical study of the citation distribution. *European Physical Journal B*, *4*, 131–134.

Redner, S. (2005). Citation statistics from more than a century of physical review. *Physics Today*, *58*, 49. See http://arxiv.org/abs/physics/0407137

Xie, H., Walker, D., Yan, K.-K, Maslov, S., & Redner, S. (in preparation). Ranking scientific publications using a simple model of network traffic.

Available online at www.sciencedirect.com
ScienceDirect
Journal of Informetrics 1 (2007) 83–91
Journal of INFORMETRICS An International Journal
www.elsevier.com/locate/joi

Gatekeepers of science—Effects of external reviewers' attributes on the assessments of fellowship applications

Lutz Bornmann [a,*], Hans-Dieter Daniel [a,b]

[a] *ETH Zurich, Professorship for Social Psychology and Research on Higher Education, Switzerland*
[b] *University of Zurich, Evaluation Office, ETH Zurich*

Received 18 May 2006; received in revised form 20 September 2006; accepted 21 September 2006

Abstract

Aim: The scientific norm of universalism prescribes that external reviewers recommend the allocation of awards to young scientists solely on the basis of their scientific achievement. Since the evaluation of grants utilizes scientists with different personal attributes, it is natural to ask whether the norm of universalism reflects the actual evaluation practice.
Subjects and methods: We investigated the influence of three attributes of external reviewers on their ratings in the selection procedure followed by the Boehringer Ingelheim Fonds (B.I.F.) for awarding long-term fellowships to doctoral and post-doctoral researchers in biomedicine: (i) number of applications assessed in the past for the B.I.F. (reviewers' evaluation experience), (ii) the reviewers' country of residence and (iii) the reviewers' gender. To analyze the reviewers' ratings (1: award; 2: maybe award; 3: no award) in an ordinal regression model (ORM) the following were considered in addition to the three attributes: (i) the scientific achievements of the fellowship applicants, (ii) interaction effects between reviewers' and applicants' attributes and (iii) judgmental tendencies of reviewers.
Results: The results of the model estimations show no significant effect of the reviewers' attributes on the evaluation of B.I.F. fellowship applications. The ratings of the external reviewers are mainly determined by the applicants' scientific achievement prior to application.
Conclusions: The results suggest that the external reviewers of the B.I.F. indeed achieved the foundation's goal of recommending applicants with higher scientific achievement for fellowships and of recommending those with lower scientific achievement for rejection.

Keywords: Peer review; Particularism; Universalism; Country of residence; Gender; Evaluation experience; Judgmental tendencies

1. Introduction

Since the classical studies of Cole (1992), Cole and Cole (1981) on the peer review process of the National Science Foundation (NSF, Arlington, VA, USA), many studies have been conducted that were aimed at examining whether the peer review procedure actually lived up to the ideal norm of universalism (Owen, 1982; Pruthi, Jain, Wahid, Mehra, & Nabi, 1997; Ross, 1980; Sharp, 1990). The results of these studies suggest that the evaluation of new work is influenced

* Corresponding author. Tel.: +41 44 632 48 25; fax: +41 44 632 12 83.
E-mail address: bornmann@gess.ethz.ch (L. Bornmann).

1751-1577/$ – see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.joi.2006.09.005

by a complex interaction between (i) universalistic factors, such as scientific merit, and (ii) scientific and non-scientific particularistic factors, such as gender. Based on these findings, Cole (1992) assumes that there is no way to objectively evaluate new scientific work. The particularistic factors that were examined in connection with the grant peer review process primarily were attributes of applicants. There are hardly any studies available on the effect of specific reviewers' attributes that may possibly contribute to bias in peer review.

We investigated the peer review procedure of the Boehringer Ingelheim Fonds (B.I.F.) – a foundation for the promotion of basic research in biomedicine (Fröhlich, 2001) – for awarding long-term fellowships to doctoral and post-doctoral researchers. In evaluating the selection process at the B.I.F. we examined the extent to which particularism has a decisive influence on judgments. As the first evaluation step, we examined whether applicants' sex, nationality, major field of study and institutional affiliation could have influenced the fellowship award decisions. For post-doctoral fellowships, no statistically significant influence of any of these variables could be observed. For doctoral fellowships, we found evidence of institutional, major field of study and gender bias, but not of a nationality bias. Furthermore, we analysed the extent to which the foundation's Board of Trustees' practice of reviewing the applications in alphabetic order when making final selection decisions influences the decisions that are made. A statistically significant influence could be observed, but the magnitude of the effect was small. The results of the first evaluation step were reported earlier (Bornmann & Daniel, 2005a,b,c, 2006).

Since the secretariat of the B.I.F. obtains the expert opinions of external reviewers for the fellowship award decisions of the Board of Trustees, we have in a second evaluation step for the present study determined to what extent the decisions of the external reviewers regarding fellowship applications are influenced by particularistic attributes (scientific and non-scientific) of the external reviewers themselves:

1. *Number of applications assessed in the past for the B.I.F.* (*reviewer's evaluation experience*): A large number of external reviewers of the B.I.F. rated more than one application during the investigation period. We assume that external reviewers, who have frequently evaluated applications for the B.I.F., are more experienced in dealing with the selection criteria of the B.I.F. than reviewers who have rarely communicated their expert opinion to the B.I.F. during the investigation period. Therefore, it is important to examine whether the number of applications reviewed by each external reviewer (e.g., the evaluation experience) gained by each external reviewer, has any effect on reviewers' recommendations (Jayasinghe, Marsh, & Bond, 2001; Kliewer, Freed, DeLong, Pickhardt, & Provenzale, 2005; Moed, 2005).
2. *Reviewer's gender*: The question of whether the reviewer's gender has an effect on recommendations is of general interest. Of particular interest is whether female or male reviewers give systematically more favourable or unfavourable recommendations to female or male fellowship applicants (Jayasinghe et al., 2001; Sonnert, 1995). According to the 'matching hypothesis', "external reviewers give higher ratings to applicants who are more similar to them on important background characteristics" (Jayasinghe, 2003, p. 7).
3. *Reviewer's country of residence*: Many external reviewers with country of residence outside Germany are asked by the B.I.F. administrative office to review B.I.F. fellowship applications. An important issue is whether recommendations by external reviewers from other countries differ from recommendations given by domestic external reviewers (Jayasinghe et al., 2001). In examining this issue the validity of the 'matching hypothesis' should be evaluated (Daniel, 1993).

2. The selection procedure of the Boehringer Ingelheim Fonds (Fröhlich, 2001, 2004)

Junior scientists submit their fellowship applications to the administrative office (secretariat) of the foundation. The office forwards each application to an independent external reviewer (one reviewer for each application). When making their decision about an application, the external reviewer should seek answers to the following questions.

Applicant's achievements: What personal qualities has the applicant demonstrated during his training: talent and inquisitiveness, versatility and creativity, determination and motivation, diligence and perseverance? What are his weaknesses? Is he capable of independent research? Does he have a wide variety of techniques at his disposal? Have his results already been presented in appropriate scientific publications?
Originality of the research project: Is the project imaginative and promising? Is it likely to yield new insights or is it simply an industrious but uninspiring piece of work? Is the current status of knowledge correctly described and

adequately documented? Is the applicant's own groundwork thorough? Are the methods of investigation sophisticated and encouraging? Is the work schedule logical and realistic?
Standard of the laboratories: Has the applicant shown mobility or has he, for good reasons, been rather settled? Is it a suitable time to change the group? Do the laboratories in which he is working or plans to work have first-class equipment and an international reputation? Does the intended research project stand out sufficiently against the current investigations of the group or will the applicant simply be a welcome addition to the present team?

On the basis of these questions, the external reviewer assesses the applicant, the proposed research project and the institution at which the project will be conducted and in a final statement recommends approval or rejection.

In addition to the assessment by an external reviewer, a member of the foundation's staff interviews the applicant personally. Finally, the application, together with the external review and the staff report on the personal interview, is submitted to the B.I.F. Board of Trustees. Seven internationally renowned scientists make up the Board. At each of the three annual Board meetings, the scientists decide on applications.

3. Method

Since the external reviewers themselves did not use a rating scale, two experts of the International Centre for Higher Education Research Kassel (INCHER-Kassel, Germany) independently rated all reviews afterwards according to the rating scale: 1: award; 2: possible award; 3: no award. The reliability of the two experts' ratings is very high (kappa coefficient = 0.96).

To identify the effect of every single attribute of the reviewers (number of applications assessed in the past, gender, country of residence) on reviewers' ratings for doctoral and post-doctoral applications we used multiple ordinal regression models (ORM) (StataCorp, 2005; Long & Freese, 2006, Chapter 5). Ordinal responses arise when the variable is coded as a consecutive integer from 1 to the number of categories that can be ordered. As with the binary regression model, the ORM is non-linear, and the magnitude of the change in the outcome probability for a given change in one of the independent variables depends on the levels of all of the independent variables (Long, 1997).

Normally, when examining the influence of (scientific and non-scientific) particularistic attributes on reviewers' ratings it is impossible to establish unambiguously, whether applications from a particular group of young scientists receives more favourable ratings due to these attributes, or if the more favourable ratings are simply a consequence of the applicants' scientific merit. In other words, the influence of reviewers' attributes upon their ratings may in fact be due to universalistic factors such as differences in applicants' publication records. As the B.I.F. had information on the applicants' scientific achievements up to the date of their fellowship applications, we could therefore include not only the particularistic factors, but also the applicants' achievements as independent variables in the ORM. This proceeding in the statistical analysis of particularism is called the "control variable approach" (Cole & Fiorentine, 1991, p. 216).

All in all, 1003 applications for a doctoral and 326 for a post-doctoral fellowship received by the foundation between 1985 and 2000 and the corresponding external reviews could be included in the model estimation. Even if the whole data set of the B.I.F. evaluation study consists of 2697 applications (Bornmann & Daniel, 2005a), the ORM had to be calculated with reduced sample sizes. Only those cases could be included in the statistical analyses that had no missing values for the variables entered into the model. As a result, 51% ($n = 1003$) of the applicants for doctoral fellowships and 44% ($n = 326$) of the applicants for post-doctoral fellowships could be included. Although it is possible to include cases with missing data in the analysis using imputation methods (Mander & Clayton, 1999; Rubin & Schenker, 1986) such as provided by the statistical package Stata (StataCorp, 2005), the parameter estimates fluctuate depending on the imputation method or – in some imputation methods – according to the number of imputations performed (Schafer, 2000). Because the parameters estimated in this way vary highly and in part can hardly be replicated, no imputation methods were used for the model estimates.

4. Results

Of the 1003 applications for a doctoral and 326 applications for a post-doctoral fellowship the reviewers recommended awarding foundation fellowships to 62% of the applications for a doctoral ($n = 621$) and 58% of the applications for a post-doctoral ($n = 190$) fellowship. He or she recommended a "possible award" for 18% of the doctoral ($n = 180$) and for 20% of the post-doctoral ($n = 65$) applications and "no award" for 20% of the doctoral ($n = 202$) and 22% of

Table 1
Description of the independent variables (universalistic and particularistic factors)

Independent variable	Applicants for doctoral fellowships ($n = 1003$)		Applicants for post-doctoral fellowships ($n = 326$)	
	Values	Mean value or percent of value '1'	Values	Mean value or percent of value '1'
Year of Board of Trustees' meeting	1985 → 2000	1994.8	1990 → 1995	1993
Applicants' scientific achievement (universalistic factors)				
Applicant's age at the time of the final degree	22 → 34	25.9	–	–
Final grade (0.88 = highest grade)	0.88 → 3.2	1.3	–	–
Applicant's age at the time of receiving Ph.D.	–	–	23 → 36	28.6
h-index of the applicant	–	–	0 → 13	2.8
Number of journal articles published by applicant at the time of application	–	–	0 → 23	3.7
Reviewers' attributes (particularistic factors)				
Number of applications evaluated in the past for the B.I.F. (reviewers' evaluation experience)	1 → 15	3	1 → 12	2.9
Reviewer's gender				
Male reviewer, male applicant (=1, 0 = other combinations)	0 → 1	54%	0 → 1	58%
Male reviewer, female applicant (=1, 0 = other combinations)	0 → 1	38%	0 → 1	36%
Female reviewer, male applicant (=1, 0 = other combinations)	0 → 1	5%	0 → 1	3%
Female reviewer, female applicant (=1, 0 = other combinations, reference category)	0 → 1	3%	0 → 1	3%
Reviewer's nationality (1 = German, 0 = foreign)	0 → 1	95%	–	–
Reviewer's nationality				
German reviewer, German applicant (=1, 0 = other combinations)	–	–	0 → 1	63%
German reviewer, foreign applicant (=1, 0 = other combinations)	–	–	0 → 1	32%
Foreign reviewer, foreign applicant (=1, 0 = other combinations)	–	–	0 → 1	4%
Foreign reviewer, German applicant (=1, 0 = other combinations, reference category)	–	–	0 → 1	1%

the post-doctoral ($n = 71$) applications. The relationships between reviewers' ratings and decisions of the B.I.F. Board of Trustees (0 = approved, 1 = rejected) are *Cramer's* $V = 0.36$ (applications for a doctoral fellowship) and *Cramer's* $V = 0.27$ (applications for a post-doctoral fellowship).

Table 1 lists the independent variables (universalistic and particularistic factors) that were included in the ORMs for doctoral and post-doctoral applicants. With regard to the applicants' scientific achievements (universalistic factors), it was possible to include the age at the time of the final degree (range: 22–34, mean value: 25.9) and the final grade (range: 0.88–3.2, mean value: 1.3) for the doctoral applicants. The ORM for post-doctoral applicants included the age at the time of receiving the Ph.D. (range: 23–36, mean value: 28.6).

Since applicants for a B.I.F. doctoral fellowship have rarely published their own work prior to applying for a fellowship (Fröhlich, 2004), bibliometric measures in the analysis could only be used for post-doctoral applicants: (i) the number of journal articles (full length articles, letters, notes, communications and reviews) published by the time of application (range: 0–23, mean value: 3.7) and (ii) the applicant's *h*-index (Bornmann & Daniel, 2005d). Hirsch (2005) has proposed the *h*-index as a single-number criterion to evaluate the scientific output of a researcher (Ball, 2005). The *h*-index depends on both the number of an applicant's articles, and their impact on his or her peers: "A scientist has index *h* if *h* of his or her Np papers have at least *h* citations each and the other (Np − *h*) papers have $\leq h$ citations each" (Hirsch, 2005, p. 16569). The *h*-index does not measure the total impact of a scientist, but the breadth of the highly cited research.

We determined the citation counts by using the online database Science Citation Index (SCI; provided by Thomson Scientific, Philadelphia, PA, USA). The post-doctoral applicants have on average an *h*-index of 2.8 (see Table 1), i.e., they have written approximately three articles that have each had at least three citations from year of publication to the end of 2001. The applicant's *h*-indices range from 0 to 13.

The following attributes of the reviewers are included in the analysis as particularistic factors.

Number of applications assessed in the past for the B.I.F. (*reviewer's evaluation experience*): For each application it was determined how many other applications the corresponding external reviewer had previously evaluated for the B.I.F. A total of 539 applications (41%) were evaluated by an external reviewer who had not previously given his expert opinion to the B.I.F. For 244 applications (18%) the administrative office of the B.I.F. selected a reviewer, who had previously evaluated one other application. In 12% ($n = 158$) or 8% ($n = 107$) of the applications the reviewer already had acquired more extensive experience with the B.I.F. selection procedure by evaluating two or three other applications. A total of 281 applications (21%) were submitted to a reviewer, who had comprehensive experience in evaluating applications (between 4 and 15 evaluations). Table 1 shows that both for the doctoral and post-doctoral fellowships the corresponding reviewer on average had evaluated two other applications.
Reviewer's gender: Concerning the B.I.F. reviewer's gender, we have the opportunity to examine interaction effects between the attribute of the reviewer and the attribute of the applicant (Jayasinghe et al., 2001; Sonnert, 1995). Table 1 shows that 54% of the applications for a doctoral and 58% of the applications for a post-doctoral fellowship were submitted by a male applicant and then evaluated by a male reviewer. In 36% (post-doctoral) or 38% (doctoral) of the applications the administrative office of the B.I.F. selected a male reviewer for the application of a female applicant. In both groups (doctoral and post-doctoral applications) less than 10% of the applications were evaluated in the combinations "female reviewer/male applicant" or "male reviewer/female applicant".
Reviewer's country of residence: The archived data of the administrative office of the B.I.F. indicates the country of residence for each external reviewer, who has evaluated an application between 1985 and 2000. Expert opinions were mainly obtained from German reviewers ($n = 1256$); rarely ($n = 73$) was a foreign reviewer used (mostly from other German-speaking countries): Switzerland = 41, Austria = 18, France = 5, USA = 2, and Israel = 1 (for six applications the reviewer's country abroad is unknown).

Since the number of applications associated with the five foreign countries of the reviewers is relatively small, applications reviewed by foreign reviewers were grouped together for the ORM. Because we know the country of residence of the external reviewer and the nationality of the applicant, we have the opportunity to determine the interaction between the two variables for post-doctoral applications. Table 1 shows that 63% of the applications were submitted by a German applicant and 32% of the applications were submitted by a foreign applicant, which were then evaluated by a German reviewer. Six percent of the applications were evaluated in the combinations "foreign reviewer/foreign applicant" (4%) and "foreign reviewer/German applicant" (2%). Since the combination "foreign reviewer/foreign applicant" occurs only once in the applications for a doctoral fellowship, we were only able to distinguish between evaluations by German (95%) and foreign (5%) reviewers (see Table 1).

The total 1003 doctoral and 326 post-doctoral applications that were included in the ORMs were evaluated by a total of 642 external reviewers. Accordingly, between 1985 und 2000 each reviewer on average wrote two evaluations (range: between 1 and 13 evaluations). Previous studies (Daniel, 1993; Opthof, Coronel, & Janse, 2002; Siegelman, 1991) have consistently shown that systematic tendencies of external reviewers exist in the framing of judgments toward favourable or unfavourable evaluations during peer review. For example, in the journal *Angewandte Chemie* (Daniel, 1993) the mean ratings of eight reviewers, who had received 10 or more manuscripts during 1984, were compared. It was shown that some reviewers could be classified as belonging to the category of "assassins" and some to the category of "zealots".

Likewise, B.I.F. reviewers, who had evaluated 10 or more applications between 1985 and 2000 (8 out of a total of 624 reviewers), exhibit systematic tendencies in the framing of judgments. The median ratings of the eight reviewers ranged from 1 to 3. The result of a Kruskal–Wallis test (Kruskal & Wallis, 1952) shows that the medians of the eight reviewers exhibit a statistically significant difference, χ^2 (7, $n = 106$) = 17.1, $p < 0.05$ (Fröhlich, 2004, p. 228). This result indicates that systematic tendencies in the framing of reviewers' judgments (Schafer, 2000) exist in the evaluation of B.I.F. applications—consistent with the findings concerning journal peer review (Daniel, 1993; Opthof et al., 2002; Siegelman, 1991). For the ORM this means that our data set violates the assumption of independent ratings. As it is not the reviewers themselves, but instead the rating for each application that formed the unit of analysis, each reviewer (or the framing of his or her judgment) that has evaluated more than one application enters into the calculation of expected values several times. Using the cluster-option provided in the statistical package Stata (StataCorp, 2005; Long & Freese, 2006, pp. 85–87), the dependency of the ratings can be taken into account. The option specifies that

Table 2
Ordinal regression model (ORM) predicting external reviewers' ratings of applications for a doctoral fellowship ($n = 1003$)

Independent variable	Coefficient	Robust standard error	p-Value
Year of Board of Trustees' meeting	0.02	0.02	0.406
Applicant's scientific achievement (universalistic factors)			
Applicant's age at the time of the final degree	0.09	0.04	0.038
Final grade (0.88 = highest grade)	0.01	0.00	0.000
Reviewer's attributes (particularistic factors)			
Number of applications evaluated in the past for the B.I.F. (reviewer's evaluation experience)	0.04	0.03	0.106
Reviewer's gender			
Male reviewer, male applicant (=1, 0 = other combinations)	−0.64	0.35	0.070
Male reviewer, female applicant (=1, 0 = other combinations)	−0.52	0.35	0.140
Female reviewer, male applicant (=1, 0 = other combinations)	−0.41	0.47	0.384
Reviewer's nationality (1 = German, 0 = foreign)	0.49	0.40	0.214

the reviewers' ratings are independent across the clusters (here the clusters are the external reviewers), but are not necessarily independent within clusters.

Tables 2 and 3 show the results of the ORMs predicting the external reviewers' ratings of applications for a doctoral (Table 2) and post-doctoral (Table 3) fellowship based on universalistic factors (applicant's scientific achievement) and particularistic factors (attributes of the reviewers). In both models, statistically significant effects for applicant's scientific achievement could be found. In the applications for a doctoral fellowship the applicant's age at the time of the final degree and the final grade are statistically significant and the sign of the coefficients are in the expected direction. The calculation of percent change coefficients for reviewers' ratings according to the ORM estimation (Long & Freese, 2006, pp. 218–220) show: (i) with each additional year taken to obtain the final degree the odds of obtaining more unfavourable ratings increased by 9%, if all other variables are kept constant. (ii) With each one-tenth that the final grade decreases the odds of obtaining more favourable ratings increased by 10%.

Table 3 shows that among the applications for a post-doctoral fellowship the effect of the applicant's age at the time of receiving the Ph.D. on the ratings is statistically non-significant. In contrast, both bibliometric measures were

Table 3
Ordinal regression model (ORM) predicting external reviewers' ratings of applications for a post-doctoral fellowship ($n = 326$)

Independent variable	Coefficient	Robust standard error	p-Value
Year of Board of Trustees' meeting	0.03	0.08	0.718
Applicant's scientific achievement (universalistic factors)			
Applicant's age at the time of receiving Ph.D.	−0.00	0.06	0.968
h-index of the applicant	−0.40	0.12	0.001
Number of journal articles published by applicant at the time of application	0.24	0.08	0.005
Reviewer's attributes (particularistic factors)			
Number of applications evaluated in the past for the B.I.F. (reviewer's evaluation experience)	0.05	0.05	0.338
Reviewer's gender			
Male reviewer, male applicant (=1, 0 = other combinations)	0.18	0.91	0.841
Male reviewer, female applicant (=1, 0 = other combinations)	0.58	0.91	0.524
Female reviewer, male applicant (=1, 0 = other combinations)	0.69	1.14	0.545
Reviewer's nationality			
German reviewer, German applicant (=1, 0 = other combinations)	−0.04	0.58	0.948
German reviewer, foreign applicant (=1, 0 = other combinations)	−0.31	0.59	0.604
Foreign reviewer, German applicant (=1, 0 = other combinations)	−0.59	1.02	0.561

statistically significant. The odds of getting more favourable ratings increase by 33% for every unit increase in the *h*-index, while keeping all other variables constant. An unexpected result is shown for the coefficient of the variable "number of journal articles published by applicant at the time of application". For each additional article the odds of getting more favourable ratings *decrease* by 27%. Accordingly, many articles *increase* the chance that the application will be rejected.

With regard to the influence of particularistic factors (three attributes of reviewers) on reviewers' ratings, both the applications for a doctoral (Table 2) as well as a post-doctoral (Table 3) fellowship did not experience any statistically significant effects from the (i) number of applications assessed in the past for the B.I.F. (reviewer's evaluation experience), nor the various combinations of (ii) reviewer's and applicant's gender and the (iii) reviewer's country of residence and applicants' nationality. This result suggests that during the B.I.F. peer review performed in accordance with the criteria provided to the external reviewers for their evaluation the ratings of the reviewers are based on the applicants' scientific achievements and that the ratings are hardly influenced by particularistic factors introduced through certain attributes of the reviewers.

5. Discussion

In this study, we have used ordinal regression models (ORMs) to examine the influence of particularistic and universalistic factors on the assessment process in the sciences. Using the data of the B.I.F., we have checked the influence exerted by three attributes of external reviewers and applicants' scientific achievements on the evaluation of fellowship applications that were submitted to the B.I.F. between 1985 and 2000. In the following, we would like to discuss the results of this study in light of the background of the findings made by other studies:

1. *Number of applications evaluated in the past for the B.I.F.* (*reviewer's evaluation experience*): In a comprehensive study the Australian Research Council (ARC, Canberra) evaluated the funding of Australian university research across all disciplines (Jayasinghe et al., 2001). With regard to the number of applications reviewed in the past by an ARC external reviewer, the results show that the reviewers' ratings tend to become more unfavourable the more frequently reviewers had evaluated applications for the ARC (i.e., the more experience they had with the peer review process of the ARC). Even if this tendency of the reviewers' ratings likewise can be detected in the B.I.F. peer review process of this study, the influence of the number of prior evaluations on the reviewers' ratings in the ORMs is shown to be statistically non-significant.
2. *Reviewer's gender*: According to United States General Accounting Office (1994) the NSF and the National Endowment for the Humanities (NEH, Pennsylvania, NW, Washington, DC, USA) have policies to promote reviewer selection that is balanced in terms of race, gender, and religion. Nevertheless, women external reviewers are underrepresented in both agencies: "Only 6 percent of NSF reviewers were women, compared to 21 percent at NEH" (United States General Accounting Office, 1994, p. 39). Also at the B.I.F. an application is evaluated by a female reviewer clearly less often (7% of applications) than by a male reviewer (93% of applications).

   An experimental study (Sonnert, 1995) found that grant submissions by women biologists received even better average ratings than the grant submissions by men (mean rating: 3.67 versus 3.27; $p = 0.0496$). If the gender of the evaluators in the data analysis is considered, the following result can be seen: "Women raters, as a group, gave the biologists substantially better quality ratings than men raters did, but they gave higher scores equally to women and men biologists. Thus, no biases arose from particular combinations of the gender of evaluators and those evaluated" (p. 47). A comparable result is obtained in a study concerning the grants peer review of the ARC (Jayasinghe et al., 2001): "Main effects due to the gender of the first researcher and the gender of the external reviewer and their interaction were all statistically non-significant" (p. 353). Also in this study, we examined the ratings of the external reviewers with regard to an interaction effect between the reviewers' gender and the applicants' gender. In agreement with the results of both previous studies (Jayasinghe et al., 2001; Sonnert, 1995), the differences in the ratings between the four groups with different gender combinations are statistically non-significant.
3. *Reviewer's country of residence*: The study concerning the ARC peer review (Jayasinghe et al., 2001) also examined the question of whether ratings given by Australian external reviewers differ from ratings given by external reviewers from other countries. The result shows that Australian external reviewers gave significantly lower ratings than did non-Australian reviewers, particularly those from North America. Possible interaction effects between the applicants' and reviewers' country of residence were not investigated. Since the B.I.F. peer review not only provides

information on the reviewer's country of residence, but also the nationality of the fellowship applicants, we were able to evaluate interaction effects (at least for post-doctoral applicants) in this study. Both while considering the interaction effects (post-doctoral applications) and without considering these effects (doctoral applicants), our results show no statistically significant influence of the reviewer's country of residence on the ratings.

In addition to particularistic factors the ORMs included measures of the applicant's scientific achievement to control the universalistic factors in the statistical analysis. The results of the model estimations for doctoral and post-doctoral applicants show that of the five included scientific achievement measures four exert a statistically significant influence on the external reviewers' ratings. In the applications for a doctoral fellowship the ratings are determined by (i) the applicant's age at the time of obtaining the final degree and (ii) his or her final grade; in the applications for a post-doctoral fellowship the main factors are (iii) the number of journal articles published at the time of application and (iv) the impact of these articles (*h*-index). While the applicant's age (i), the final grade (ii) and the *h*-index (iv) are consistent with the expected influence on the ratings, the number of articles (iii) yields an unexpected result: each additional article *reduces* the chance for a favourable evaluation. This unexpected result only can be explained by considering the influence of the *h*-index on the ratings. The size of the *h*-index is dependent primarily on an applicant's number of articles that have achieved substantial impact (citations by scientific colleagues). Accordingly, the result of the ORMs suggest that only those articles to which reviewers attribute a substantial impact lead to more favourable ratings. If this impact is not given in the reviewer's opinion, each additional article *reduces* the chance for a favourable rating.

These results suggest that the external reviewers of the B.I.F. indeed achieved the foundation's goal of recommending applicants with higher scientific achievement for fellowships and of recommending those with lower scientific achievement for rejection. Similar findings (Chapman & McCauley, 1994) were reported for quality ratings of graduate fellows funded by the National Science Foundation (NSF, Arlington, VA, USA). Results of a study on the committee peer review of a National Research Council in a smaller Western-European country (Moed, 2005, pp. 247–257) show that the median citation impact of applicants, who were rated excellent by their peers, is higher than that of all other applicants. Similar results have been reported for selection decisions in the journal peer review process. Based on the mean citation rates for accepted and rejected manuscripts that were nevertheless published elsewhere, the decisions made by the editors of the *Journal of Clinical Investigation* (Wilson, 1978), *British Medical Journal* (Lock, 1985) and *Angewandte Chemie* (Daniel, 1993) reflect a high degree of validity.

Overall, the results of this study suggest that the B.I.F. peer review is valid and hardly influenced by gender, nationality and the number of prior evaluations performed by the reviewer (three particularistic factors). Even if the influence of certain reviewers' attributes on the ratings has been shown to be of little significance, our result from a Kruskal–Wallis test prior to the ORMs shows that the B.I.F. peer review is not a purely objective process: some external reviewers belong in strict ("assassins") and some in lenient ("zealots") judgment categories.

Acknowledgements

The research was supported by the International Centre for Higher Education Research Kassel (INCHER-Kassel) and by the University of Zurich.

References

Ball, P. (2005). Index aims for fair ranking of scientists. *Nature*, *436*(7053), 900.

Bornmann, L., & Daniel, H.-D. (2005a). Selection of research fellowship recipients by committee peer review. Analysis of reliability, fairness and predictive validity of Board of Trustees' decisions. *Scientometrics*, *63*(2), 297–320.

Bornmann, L., & Daniel, H.-D. (2005b). Committee peer review at an international research foundation: Predictive validity and fairness of selection decisions on post-graduate fellowship applications. *Research Evaluation*, *14*(1), 15–20.

Bornmann, L., & Daniel, H.-D. (2005c). Criteria used by a peer review committee for selection of research fellows—A boolean probit analysis. *International Journal of Selection and Assessment*, *13*(4), 296–303.

Bornmann, L., & Daniel, H.-D. (2005d). Does the *h*-index for ranking of scientists really work? *Scientometrics*, *65*(3), 391–392.

Bornmann, L., & Daniel, H.-D. (2006). Selecting scientific excellence through committee peer review—A citation analysis of publications previously published to approval or rejection of post-doctoral research fellowship applicants. *Scientometrics*, *68*(3), 427–440.

Chapman, G. B., & McCauley, C. (1994). Predictive validity of quality ratings of National Science Foundation graduate fellows. *Educational and Psychological Measurement*, *54*(2), 428–438.

Cole, S. (1992). *Making science. Between nature and society*. Cambridge, MA, USA: Harvard University Press.

Cole, J. R., & Cole, S. (1981). *Peer review in the National Science Foundation. Phase two of a study*. Washington, DC, USA: National Academic Press.

Cole, S., & Fiorentine, R. (1991). Discrimination against women in science: The confusion of outcome with process. In H. Zuckerman, J. R. Cole, & J. T. Bruer (Eds.), *The outer circle. Women in the scientific community* (pp. 205–226). London, UK: W.W. Norton & Company.

Daniel, H. -D. (1993). *Guardians of science. Fairness and reliability of peer review* (pp. 2004). Weinheim, Germany: Wiley-VCH.

Fröhlich, H. (2001). It all depends on the individuals. Research promotion—A balanced system of control. *B.I.F. Futura*, *16*, 69–77.

Fröhlich, H. (2004). Pillars of wisdom—Interaction between trustees and reviewers. *B.I.F. Futura*, *19*, 227–228.

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*(46), 16569–16572.

Jayasinghe, U. W. (2003). *Peer review in the assessment and funding of research by the Australian Research Council*. Greater Western Sydney, Australia: University of Western Sydney.

Jayasinghe, U. W., Marsh, H. W., & Bond, N. (2001). Peer review in the funding of research in higher education: The Australian experience. *Educational Evaluation and Policy Analysis*, *23*(4), 343–346.

Kliewer, M. A., Freed, K. S., DeLong, D. M., Pickhardt, P. J., & Provenzale, J. M. (2005). Reviewing the reviewers: Comparison of review quality and reviewer characteristics at the American Journal of Roentgenology. *American Journal of Roentgenology*, *184*(6), 1731–1735.

Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, *47*(260), 583–621.

Lock, S. (1985). *A difficult balance: Editorial peer review in medicine*. Philadelphia, PA, USA: ISI Press.

Long, J. S. (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks, CA, USA: Sage.

Long, J. S., & Freese, J. (2006). *Regression models for categorical dependent variables using Stata (2 ed.):* College Station, TX, USA: Stata Press, Stata Corporation.

Mander, A., & Clayton, D. (1999). Hotdeck imputation. *Stata Technical Bulletin*, *51*, 16–18.

Moed, H. F. (2005). *Citation analysis in research evaluation*. Dordrecht, The Netherlands: Springer.

Opthof, T., Coronel, R., & Janse, M. J. (2002). The significance of the peer review process against the background of bias: Priority ratings of reviewers and editors and the prediction of citation, the role of geographical bias. *Cardiovascular Research*, *56*(3), 339–346.

Owen, R. (1982). Reader bias. *Journal of the American Medical Association*, *247*(18), 2533–2534.

Pruthi, S., Jain, A., Wahid, A., Mehra, K., & Nabi, S. A. (1997). Scientific community and peer review system—A case study of a central government funding scheme in India. *Journal of Scientific and Industrial Research*, *56*(7), 398–407.

Ross, P. F. (1980). *The sciences' self-management: Manuscript refereeing, peer review and goals in science*. Massachusetts, MA, USA: The Ross Company, Todd Pond.

Rubin, D. B., & Schenker, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association*, *81*(394), 366–374.

Schafer, J. L. (2000). *Analysis of incomplete multivariate data by simulation*. London, UK: Chapman and Hall.

Sharp, D. W. (1990). What can and should be done to reduce publication bias—The perspective of an editor. *Journal of the American Medical Association*, *263*(10), 1390–1391.

Siegelman, S. S. (1991). Assassins and zealots—Variations in peer review. Special report. *Radiology*, *178*(3), 637–642.

Sonnert, G. (1995). What makes a good scientist? Determinants of peer evaluation among biologists. *Social Studies of Science*, *25*, 35–55.

StataCorp (2005). *Stata statistical software: Release 9*. College Station, TX, USA: StataCorp LP.

United States General Accounting Office (1994). Peer review: Reforms needed to ensure fairness in federal agency grant selection. Washington, DC, USA: United States General Accounting Office.

Wilson, J. D. (1978). Peer review and publication. *Journal of Clinical Investigation*, *61*(4), 1697–1701.

Available online at www.sciencedirect.com
ScienceDirect
Journal of Informetrics 1 (2007) 16–25
Journal of INFORMETRICS An International Journal
www.elsevier.com/locate/joi

Hirsch's h-index: A stochastic model

Quentin L. Burrell *

Isle of Man International Business School, The Nunnery, Old Castletown Road, Douglas, Isle of Man IM2 1QB, via United Kingdom

Received 29 June 2006; received in revised form 27 July 2006; accepted 28 July 2006

Abstract

We propose a simple stochastic model for an author's production/citation process in order to investigate the recently proposed h-index for measuring an author's research output and its impact. The parametric model distinguishes between an author's publication process and the subsequent citation processes of the published papers. This allows us to investigate different scenarios such as varying the production/publication rates and citation rates as well as the researcher's career length. We are able to draw tentative results regarding the dependence of Hirsch's h-index on each of these fundamental parameters. We conjecture that the h-index is, according to this model, (approximately) linear in career length, log publication rate and log citation rate, at least for moderate citation rates.

Keywords: Hirsch h-index; Stochastic model; Informetric process

1. Introduction

Hirsch (2005) proposed the h-index, a single number to measure an individual's research output and its impact. In the original (preprint) version, his definition states (essentially) "*A scientist has index h if h of his/her papers have at least h citations each and the rest have fewer than h citations each*". Since its first publication, the index has received much attention, both in the popular domain, as in Ball (2005), and in the academic literature. In the latter we can distinguish straightforward attempts to apply the index, as in Bornmann and Daniel (2005), Rousseau (2006) and Cronin and Meho (2006); those seeking to modify the index or extend its range of application, as in Braun, Glänzel, and Schubert, 2005, and Egghe (2006); those that relate the h-index to the "traditional" bibliometric evaluation measures, such as Glänzel (2006b) and Van Raan (2006), and those seeking to give some sort of mathematical model for the index, as in Egghe (in press), Egghe and Rousseau (in press), Glänzel (2006a), as well as the original paper of Hirsch (2005). In this paper we use an established informetric way of modelling the production/citation process to seek to give some insight to the index.

2. The publication-citation model

Consider an author whose publishing career begins at time zero and we then wish to model the numbers of citations to each of his/her publications by time T (the present). (Already, we have a small problem as to how we define "time

* Tel.: +44 1624693706; fax: +44 1624665095.
E-mail address: q.burrell@ibs.ac.im.

1751-1577/$ – see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.joi.2006.07.001

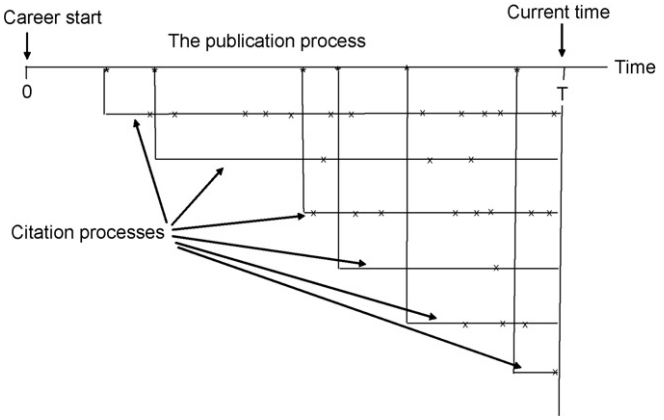


Fig. 1. Representation of the publication/citation processes.

zero". Is it the time at which the author takes up his/her post, or when the first publication is submitted, or accepted, or appears in print? In what follows, let us not worry over this detail.) In other words we are assuming that currently the individual is T time units into his/her productive career. (The unit of time will typically be 1 year, but for modelling purposes this is not important.) Thus, we assume that the author publishes papers at certain times and that these papers subsequently attract citations following their publication, where both the publication and citation accumulation processes are random. We further assume that some papers are more citable than others so that the citation rate varies between different publications. A schematic representation of the model is given in Fig. 1 where we have an author with six papers published during the period of observation. The first (earliest) of these gains 12 citations, the second 3, then 8, 1, 3 and 1 for the rest.

In order to make this general scenario analytically viable we need to be more precise, so our initial model assumes:

Assumption 1. From the start of his/her publishing career at time zero, an author publishes papers according to a Poisson process of rate θ. Thus, by time T, the number of publications Y_T has the distribution

$$P(Y_T = r) = e^{-\theta T}\frac{(\theta T)^r}{r!}, \quad r = 0, 1, 2, \ldots, \text{ and } E[Y_T] = \theta T$$

Note that the parameter θ gives the mean number of publications per unit time for this author, called the *publication rate*.

Assumption 2. Any particular publication acquires citations according to a Poisson process of rate Λ, where Λ varies from paper to paper. Here, Λ denotes the mean number of citations per unit time following publication, called the *citation rate*.

Assumption 3. The citation rate Λ for this author varies over the set of his/her publications according to a gamma distribution of index $\nu \geq 1$ and scale parameter $\alpha > 0$. Thus, the probability density function of Λ is given by

$$f_\Lambda(\lambda) = \frac{\alpha^\nu}{\Gamma(\nu)}\lambda^{\nu-1}e^{-\alpha\lambda}, \quad 0 < \lambda < \infty$$

Note that $E[\Lambda] = \nu/\alpha$ gives the overall *mean citation rate*, or the average number of citations acquired by a randomly selected paper of this author per unit time.

Remarks.

(i) The usual requirement for the gamma index ν is that it is positive. We require the restriction $\nu \geq 1$ in order to ensure convergence of certain integral expressions in the following. As this should be viewed as an exploratory model, the restriction is of no real concern.

(ii) This model has previously been suggested by Burrell (1992) to develop an idea of Rousseau (1992) and can be considered as an example of what Egghe and Rousseau (1990, p. 378) call a three-dimensional informetric study.

In Burrell (1992) the focus of interest was simply the total number of citations accumulated by the collection of publications, here we are interested in the distribution of the number of citations to the individual publications.

For any author, the "current time" T is equivalent to the time since the author's publication career began, which we take as defining time zero (for this author). For our author, let us therefore denote by X_T the number of citations achieved by a (randomly chosen) published paper by time T and by $N(n; T)$ the number of published papers receiving at least n citations by time T. Then we have:

Theorem 1. *Under the assumptions of the model, the distribution of the number of citations to a randomly chosen paper by time T is given by*

$$P(X_T = r) = \frac{\alpha}{(\nu-1)T} B\left(\frac{T}{\alpha+T}; r+1, \nu-1\right) \text{ for } r = 0, 1, 2 \ldots \tag{1}$$

where

$$B(x; a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^x y^{a-1}(1-y)^{b-1} \mathrm{d}y$$

is the cumulative distribution function of a beta distribution (of the first kind) with parameters a and b.

Proof. See the Appendix.

So far as determination of the h-index is concerned, our main result is almost a corollary of the above:

Theorem 2. *Under the assumptions of the model, the expected number of papers receiving at least n citations by time T is given by*

$$E[N(0;\ T)] = \theta T$$

$$E[N(n; T)] = \theta T \left(1 - \frac{\alpha}{(\nu-1)T} \sum_{r=0}^{n-1} B\left(\frac{T}{\alpha+T}; r+1, \nu-1\right)\right) \text{ for } n = 1, 2, 3, \ldots \tag{2}$$

Proof. The proof of this result follows from a pair of Lemmas, the proofs of which are given in the Appendix. The result for $n = 0$ is of course just the expected number of publications by time T. □

Lemma 1.

$$E[N(n; T)] = \frac{\alpha\theta}{(\nu-1)} \sum_{r=n}^{\infty} B\left(\frac{T}{\alpha+T}; r+1, \nu-1\right)$$

Lemma 2.

$$\sum_{r=0}^{\infty} B\left(\frac{T}{\alpha+T}; r+1, \nu-1\right) = \frac{(\nu-1)T}{\alpha}$$

Remarks.

(a) Note that for given parameter values and T, evaluation of the right hand side of (2) is straightforward with any statistical package including evaluation of the cumulative distribution function of the Beta distribution.

(b) Although we have adopted a notation emphasising the dependence on n and T, the expected number of citations actually depends on the model parameters θ, α and ν so perhaps we should write $E[(n; T)|\theta, \alpha, \nu]$. Although this is rather cumbersome, it simplifies when we see from (2) that

$$E[(n; T)|\theta, \alpha, \nu] = \theta E[(n; T)|1, \alpha, \nu] \tag{3}$$

This is useful in later calculations.

(c) The following are intuitively obvious, but note from (2) that the expected number of papers receiving at least n citations
  (i) is proportional to θ, the publication rate. Hence, all other things being equal, more productive authors achieve more (expected) citations.
  (ii) is increasing in T for any n.
  (iii) is decreasing in n for any T.

3. The h-index

The h-index as originally defined for an individual scientist is that number h such that h of his/her papers have at least h citations, while the rest have no more than h citations. Although Hirsch (2005) wrote in terms of papers written by scientists, there is no reason not to extend this to any academic discipline, hence we talk of authors. Also, Glänzel (2006a) has pointed out that there is a possible ambiguity in the case where an author has several papers with the same number of citations at h. (This was realised by Hirsch (2005) in the published version of his paper.)

Thus, the h-index is concerned with the distribution of the number of citations accumulated by each of an author's publications up to the current time. In our notation, Hirsch's index is given by:

Definition 1. Hirsch's h-index at time T is, for any particular author, the integer $h(T)$ satisfying

$$h(T) = \max\{n : n \leq N(n; T)\}$$

Note that this is the modified form proposed by Glänzel (2006a) to take account of possible ties. Also, this is an empirical measure, requiring observation of the actual values of $N(n; T)$. Let us remark at this stage that the upper bound for any author is his/her total number of publications.

Here we are considering a theoretical model so let us modify the above to:

Definition 2. The theoretical h-index at time T is the integer $h(T)$ satisfying

$$h(T) = \max\{n : n \leq E[N(n; T)]\}$$

Note that this is well defined since we have already remarked that $E[N(n; T)]$ decreases with n. Also, since $E[N(n; T)]$ is the expected value of a random variable and hence is not necessarily an integer it will be useful to define

$$h^*(T) = E[N(h(T); T)]$$

being the expected number of papers receiving at least $h(T)$ citations and note that necessarily $h^*(T) \geq h(T)$.

4. Exploring different scenarios

Our model involves four parameters:

(i) The author's publication rate, θ.
(ii) The gamma parameters, ν and α.
(iii) The length of the author's publishing career to the current time, T.

All four of these parameters can vary from author-to-author. So far as the publication rate is concerned, some authors are very prolific, others less so. For the gamma parameters, note that ν/α is the average citation rate for this author's papers. This average citation rate may reflect the author's stature in the field but may well also be dependent on the field as different fields have different citation practices. In addition, those working in smaller subject areas are addressing smaller audiences from whom citations are generated. Clearly the length of the publishing career also varies. We will consider various scenarios to illustrate the dependence on the individual parameters.

Table 1
Determination of the theoretical h-index: $E[N(n; 10)]$ for varying production rate, θ

n	$\theta = 2$	$\theta = 5$	$\theta = 10$
0	20.00	50.00	100.00
1	19.50	48.75	97.50
2	19.00	47.50	95.00
3	18.50	46.25	92.50
4	18.00	45.00	90.01
5	17.50	43.76	87.52
6	17.01	42.52	85.03
7	16.51	41.28	82.55
8	16.02	40.04	80.09
9	15.53	38.82	77.63
10	15.04	37.60	75.20
11	14.56	36.39	72.78
12	14.08	35.19	70.38
13	**13.60**	34.00	68.01
14	13.13	32.83	65.66
15	12.67	31.67	63.35
17	11.76	29.41	58.22
19	10.89	27.22	54.45
21	10.05	25.13	50.25
23	9.25	**23.12**	46.24
25	8.49	21.21	42.43
27	7.76	19.42	38.82
29	7.08	17.71	35.42
31	6.44	16.12	**32.24**
33	5.85	14.63	29.27

4.1. Varying the publication rate

To illustrate the determination of the h-index we consider first a trio of authors who differ only in their publication rate. Suppose that these three publish on average $\theta = 2$, 5, and 10 papers per year, respectively, and that the gamma parameters are $\alpha = 1$, $\nu = 5$ for all three so that each paper (for each author) receives five citations per year on average (after publication). Then with a publishing career of (current) length $T = 10$ years for each, calculations based on (2) lead to the results in Table 1. (Note that we have trimmed and truncated the output for the current purpose.)

Of course, since $E[N(n; T)]$ is directly proportional to θ, see (3), the columns are simply multiples of the one that could be calculated for $\theta = 1$. From the table we can read off the highlighted h-index (expected h-index) in the three cases as $h = 13$ ($h^* = 13.60$), $h = 23$ ($h^* = 23.12$) and $h = 31$ ($h^* = 32.24$), respectively. As should be expected on intuitive grounds these increase with the publication rate but note that the increase is nonlinear. This can be seen in Fig. 2(a) where the h-index for $T = 10$ is plotted against the production rate from 0 to 50, with the same gamma parameters as above. In fact we plot both h and h^* for illustration, confirming that always $h^* \geq h$.

Note that we have only considered values of θ up to 50, i.e. an author producing on average an article practically on a weekly basis, which was felt to be a reasonable upper value in practice. Fig. 2(b) shows the same data but with a logarithmic scale for the production rate and, for clarity, using only the h^* values. Here, we see almost perfect linearity. In fact, note that we have included productivity values θ as large as 500 and this does not cause any deviation from the same linear fit! (For the actual values plotted the R^2 value is greater than 0.99.) Further numerical investigation suggests that this result is robust for other values of both T and the gamma parameters. Hence, our model suggests that the h-index is approximately linear in the logarithm of the publication rate.

4.2. Varying the career length

By its very definition, an author's h-index cannot decrease in time so here we investigate its time dependence. We note that Egghe (2006) considers a time-dependent model for the h-index but his model is very different from ours

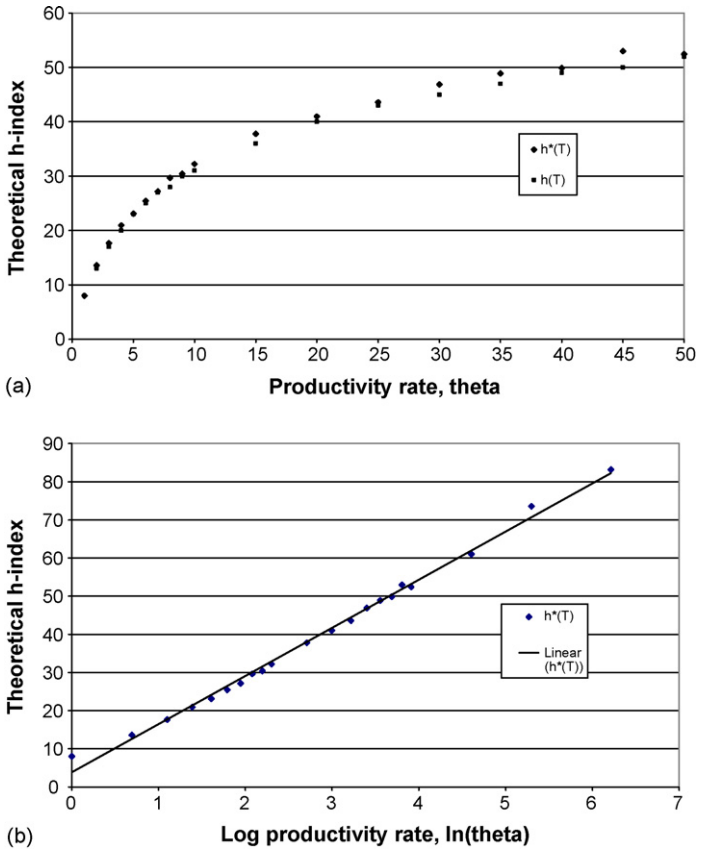


Fig. 2. (a) h-index as a function of productivity. (b) h-index as a function of log productivity.

in that he essentially supposes that an author's entire body of work is available at time zero and it is the subsequent accumulation of citations that is investigated. Similarly, Glänzel (2006a) does not model the productivity process, only the citation distribution. On the other hand, our model allows new publications to appear during the period of observation so that we are genuinely modelling the evolution of an author's publication/citation career. Given the three different author productivities and gamma parameters as in the previous section, let us consider career lengths of 5, 10 and 20 years. (Note that when we speak of career length we mean "current career length", i.e. we are thinking of an author who is currently active but whose productive career began 5, 10 or 20 years ago.) Rather than presenting the numerical analysis we adopt a graphical approach which perhaps gives greater insight into the interaction between time and productivity.

In Fig. 3 we have plotted $E[(n; T)]$ for the three different career lengths and for $\theta = 1$ in each case. Notice straight away that, as expected, $E[(n; T)]$ decreases with n for each fixed T and increases with T for any n. (Remark: Although we are looking at a function of the discrete variable n, we have plotted it as a continuous function for ease of visual interpretation.) To find the approximate h-index, at least to graphical accuracy, this is given, for any θ, by the solution of

$$n = E[N(n; T)|\theta] = \theta E[N(n; T)|1],$$

or

$$n/\theta = E[N(n; T)]$$

Hence in Fig. 3 we also plot the lines n/θ corresponding to the various θ values, allowing us to read off the (approximate) h-values. The exact values are summarised in Table 2.

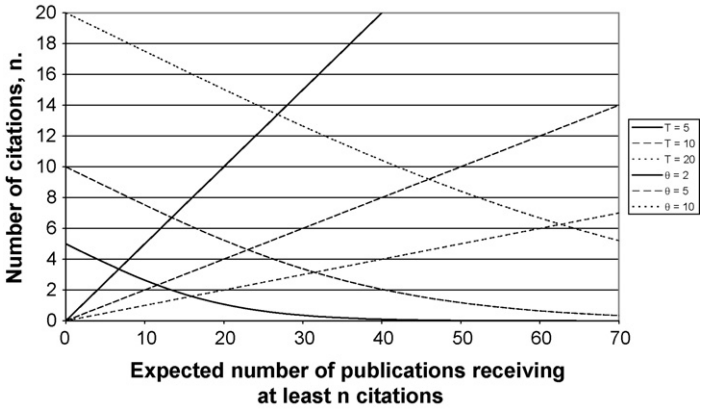


Fig. 3. h-index as a function of time and of publication productivity.

Table 2
The h-index for varying career length, T

	$T=5$ $h(h^*)$	$T=10$ $h(h^*)$	$T=20$ $h(h^*)$
$\theta=2$	6 (7.06)	13 (13.60)	26 (27.64)
$\theta=5$	11 (12.19)	23 (23.12)	45 (46.95)
$\theta=10$	15 (17.29)	31 (32.24)	62 (63.64)

The, perhaps surprising, observation from these figures is that the h-index is (almost) directly proportional to T, the length of an author's career to the current time (given the constancy of the other parameters). Further numerical investigations suggest that this approximate result is true over a wide range of parameter values. This result was in fact conjectured by Hirsch (2005). Note that Egghe (2006), using very different model assumptions found a rather more complex time dependence.

4.3. Varying the gamma parameters

In Table 3 we give the expected h-index for production rate $\theta=5$ but different career lengths and different combinations of the gamma parameters, subject to the mean citation rate ν/α over all publications being the same. For purposes of illustration we have taken this mean citation rate to be $\nu/\alpha=10$. What is surprising here is that it is not the individual parameter values that are crucial, only their ratio so that it is the overall mean citation rate which has the greatest influence, so far as the citation process is concerned, on the h-index.

In order to investigate what the relationship between the h-index and the mean citation rate might be, consider an author with $T=20$ and $\theta=5$, so the expected total number of publications is 100. To illustrate the varying citation rate, we take $\alpha=5$ and then selected values of mean citation rate between 1 and 100 so that ν varies between 5 and 500. The results for h^* are plotted in Fig. 4(a) and, with a logarithmic scale, in Fig. 4(b).

The curve in Fig. 4(a) appears to be approaching an asymptotic value of 100. This is because, just as an author's h-index has the total number of publications as its upper bound, so the theoretical index is limited by the expected

Table 3
The h-index for varying gamma parameter values, with $\nu/\alpha=10$

	$T=5$ $h(h^*)$	$T=10$ $h(h^*)$	$T=20$ $h(h^*)$
$\alpha=1$, $\nu=10$	23 (24.76)	47 (48.33)	95 (95.46)
$\alpha=2$, $\nu=20$	24 (24.79)	48 (49.53)	97 (97.97)
$\alpha=5$, $\nu=50$	24 (25.51)	49 (50.00)	98 (100.00)
$\alpha=10$, $\nu=100$	24 (25.76)	49 (50.51)	99 (100.00)
$\alpha=100$, $\nu=1000$	24 (25.98)	49 (50.95)	99 (100.90)

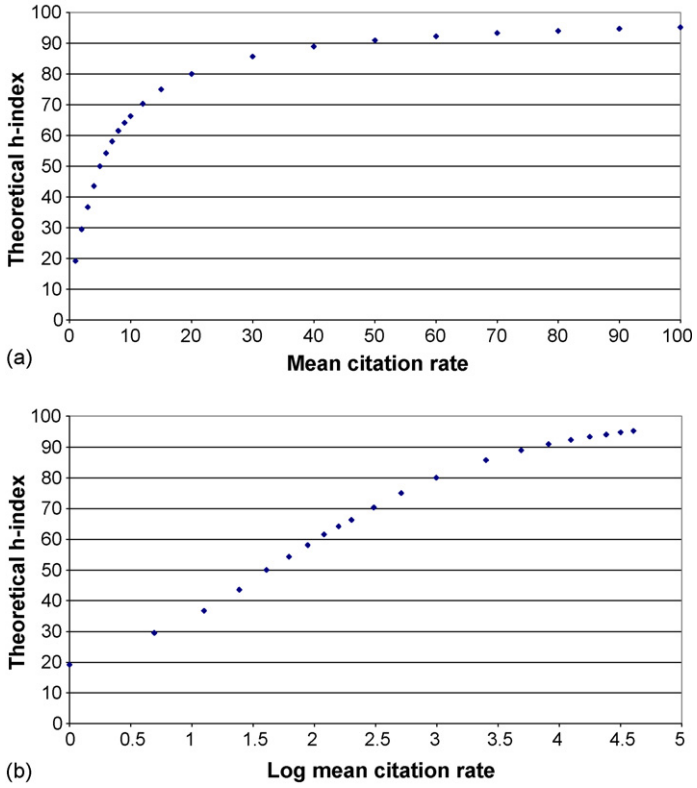


Fig. 4. (a) h-index as a function of citation rate. (b) h-index as a function of log citation rate.

number of publications, in this example 100. By the time the citation rate reaches 50, already we have $h = 90$ ($h^* = 90.96$). Turning to the log of the mean citation rate in Fig. 4(b), we have an elongated S-shape, with a levelling off at the upper end, again because of the maximal value of the h-index. However, note an approximate linearity in at least the early stages of Fig. 4(b). (For the full range of citation rates plotted, we find $R^2 = 0.970$, but clearly curvilinear. For rates up to 20, $R^2 = 0.993$ and close to linear.) Further numerical investigation suggests that, with this model, the theoretical h-index is approximately linear in the log of the citation rate, at least for "moderate" citation rates.

5. Concluding remarks

The model investigated here, based upon both the publication and citation processes being considered as Poisson processes, is perhaps the simplest genuinely stochastic model that could be considered. As such, it is worthwhile pointing out not just the general observations regarding the h-index derived from the model but also the limitations and possible modifications of the model for future work.

So far as the model is concerned, our investigations suggest that, other things being equal, Hirsch's h-index

(i) is approximately proportional to current career length, T;
(ii) is approximately a linear function of the logarithm of the author's productivity rate;
(iii) is approximately a linear function of the logarithm of the mean citation rate for the author, for moderate citation rates.

Obviously, because of their striking simplicity these observations are of interest, but note that they are based upon numerical and graphical investigations rather than analytical studies so that they should be viewed as conjectures and there is room for further theoretical work. And are there intuitively reasonable arguments to support any of them?

There is then much to be done to see if they do indeed reflect what actually happens for individual authors, i.e. does the model reflect reality? In particular, the role of the citation rate surely involves both the field in which the author works and the author's stature in that field—it would be interesting to investigate the interrelationship of these.

It is admitted that this is a simple model, based upon very specific assumptions all of which are open to challenge on intuitive grounds. For instance, is it reasonable to assume that an author's publication rate stays constant through his/her career? Also, isn't it generally accepted that a paper's citation rate varies over time with, at least eventually, a gradual decline? See Burrell (2001, 2002a, 2002b, 2003). And is the basic Poisson model the most appropriate model? What about a negative binomial process for either the productivity or, perhaps more appropriately, the citation process?

Hirsch's (2005) proposal is certainly an interesting idea for a simple index but, as a comparative measure, it would appear from our analysis that its heavy dependence on the underlying parameters means that it cannot be the *single* measure adequate "... to quantify an individual's scientific research output". We concur with Glänzel (2006a) who describes it as "a useful supplement to the bibliometric toolset" but it is certainly not a substitute.

Appendix A

Proof of Theorem 1. Let X_T denote the number of citations achieved by a published paper by time T. For a particular paper published at time $t \in [0,T]$ we have, according to the standard gamma mixture of Poisson processes (GPP) model described by Assumptions 2 and 3,

$$P(X_T = r|t) = \frac{\Gamma(r+\nu)}{r!\Gamma(\nu)}\left(\frac{\alpha}{\alpha+(T-t)}\right)^{\nu}\left(\frac{T-t}{\alpha+(T-t)}\right)^{r} \text{ for } r = 0, 1, 2, \ldots$$

Now, given the total number of publications, say $Y_T = N$, by time T, it is well known that under the assumption that these occur as a Poisson process, the successive publication times $t_1 < t_2 < \ldots < t_N$ are equivalent to the order statistics of a random sample of size N from a uniform distribution on $[0,T]$, see for instance Theorem 2.3.1, p. 67 of Ross (1996). Equivalently, the unordered times are N independent and identically distributed uniform random variables on $[0,T]$, see the Theorem on p. 75 of Stirzaker (2005). Thus, any particular publication time t can be considered to be uniformly distributed on $[0,T]$ so that

$$P(X_T = r) = E_t P(X_T = r|t) \int_0^T \frac{\Gamma(r+\nu)}{r!\Gamma(\nu)}\left(\frac{\alpha}{\alpha+(T-t)}\right)^{\nu}\left(\frac{T-t}{\alpha+(T-t)}\right)^{r}\frac{1}{T}\mathrm{d}t$$

$$= \frac{\alpha^{\nu}\Gamma(r+\nu)}{r!\Gamma(\nu)T}\int_0^T \frac{s^r}{(\alpha+s)^{r+\nu}}\mathrm{d}s \text{ where } s = T-t$$

This can easily be written in terms of a beta distribution of the second kind, see Kleiber and Kotz (2003, Chapter 6). However, if we substitute $y = s/(\alpha+s)$ so that $s = \alpha y/(1-y)$ and $\mathrm{d}s/\mathrm{d}y = \alpha/(1-y)^2$ we find for the integral

$$\int_0^T \frac{s^r}{(\alpha+s)^{r+\nu}}\mathrm{d}s = \int_0^T \left(\frac{s}{\alpha+s}\right)^{r+\nu}\frac{1}{s^{\nu}}\mathrm{d}s = \int_0^{T/(\alpha+T)} y^{r+\nu}\left(\frac{1-y}{\alpha y}\right)\frac{\alpha}{(1-y)^2}\mathrm{d}y = \int_0^{T/(\alpha+T)} \alpha^{1-\nu}y^r(1-y)^{\nu-2}\mathrm{d}y$$

Thus

$$P(X_T = r) = \frac{\alpha\Gamma(r+\nu)}{r!\Gamma(\nu)T}\int_0^{T/(\alpha+T)} y^r(1-y)^{\nu-2}\mathrm{d}y = \frac{\alpha}{(\nu-1)T}\int_0^{T/(\alpha+T)} \frac{\Gamma(r+\nu)}{\Gamma(r+1)\Gamma(\nu-1)}y^{(r+1)-1}(1-y)^{(\nu-1)-1}\mathrm{d}y$$

The integral is now just the cumulative distribution function of a beta distribution with parameters $(r+1)$ and $(\nu-1)$, provided $\nu - 1 \geq 0$ (or $\nu \geq 1$), evaluated at $T/(\alpha+T)$ so that

$$P(X_T = r) = \frac{\alpha}{(\nu-1)T}B\left(\frac{T}{\alpha+T}; r+1, \nu-1\right) \quad \square$$

Proof of Lemma 1. Noting that

$$E[N(n;T)] = E[Y_T P(X_T \geq n)] = E[Y_T]P(X_T \geq n) = \theta T \sum_{r=n}^{\infty} P(X_T = r),$$

the result follows from the above. □

Proof of Lemma 2.

$$\sum_{r=0}^{\infty} B\left(\frac{T}{\alpha+T}; r+1, \nu-1\right) = \sum_{r=0}^{\infty} \int_0^{T/(\alpha+T)} \frac{\Gamma(r+\nu)}{\Gamma(r+1)\Gamma(\nu-1)} y^r (1-y)^{\nu-2} \mathrm{d}y$$

$$= \int_0^{T/(\alpha+T)} \frac{\nu-1}{(1-y)^2} \left(\sum_{r=0}^{\infty} \frac{\Gamma(r+\nu)}{r!\Gamma(\nu)} y^r (1-y)^{\nu} \right) \mathrm{d}y$$

Now the inner summation in this integral expression on the RHS is just the total sum of the probability mass function of a NBD(y, ν) random variable and hence is equal to 1 for any y in [0,1]. Hence, we have that the original sum is equal to

$$(\nu-1) \int_0^{T/(\alpha+T)} \frac{1}{(1-y)^2} \mathrm{d}y = (\nu-1) \left[\frac{1}{(1-y)} \right]_0^{T/(\alpha+T)} = \frac{(\nu-1)T}{\alpha} \quad \square$$

References

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

Bornmann, L., & Daniel, H.-D. (2005). Does the h-index for ranking of scientists really work? *Scientometrics*, *65*(3), 391–392.

Braun, T., Glänzel, W., & Schubert, A. (2005). A Hirsch-type index for journals. *The Scientist*, *19*(22), 8–10.

Burrell, Q. L. (1992). A simple model for linked informetric processes. *Information Processing and Management*, *28*, 637–645.

Burrell, Q. L. (2001). Stochastic modelling of the first citation distribution. *Scientometrics*, *52*, 3–12.

Burrell, Q. L. (2002a). On the *n*th citation distribution and obsolescence. *Scientometrics*, *53*, 309–323.

Burrell, Q. L. (2002b). Will this paper ever be cited? *Journal of the American Society for Information Science and Technology*, *53*, 232–235.

Burrell, Q. L. (2003). Predicting future citation behavior. *Journal of the American Society for Information Science and Technology*, *54*, 372–378.

Cronin, B., & Meho, L. (2006). Using the h-index to rank influential information scientists. *Journal of the American Society for Information Science and Technology*, *57*(9), 1275–1278.

Egghe, L. (2006). An improvement of the H-index: the G-index. *ISSI Newsletter*, *2*(1), 8–9.

Egghe, L. (in press). Dynamic h-index: the Hirsch index in function of time, *Journal of the American Society for Information Science and Technology*.

Egghe, L., & Rousseau, R. (1990). *Introduction to informetrics*. Elsevier: Amsterdam.

Egghe, L., Rousseau, & R. (in press). An informetric model for the h-index. *Scientometrics*.

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

Glänzel, W. (2006b). On the opportunities and limitations of the H-index. *Science Focus*, *1*(1), 10–11 (in Chinese).

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*(46), 16569–16572. Also available as ar*Xiv:physica/0508113, accessible at http://xxx.arxiv.org/abs/physics/0508025.

Kleiber, C., & Kotz, S. (2003). *Statistical size distributions in economics and actuarial sciences*. New Jersey: Wiley.

Ross, S. (1996). *Stochastic processes* (2nd ed.). New York: John Wiley.

Rousseau, R. (1992). Concentration and diversity of availability and use in information systems. *Journal of the American Society for Information Science*, *43*, 391–395.

Rousseau, R. (2006). A case study: evolution of JASIS' Hirsch index. *Science Focus*, *1*(1), 16–17 (in Chinese).

Stirzaker, D. (2005). *Stochastic processes and models*. Oxford: Oxford University Press.

Van Raan, A. J. (2006). Comparison of the Hirsch-index with standard bibliometric indicators and with peer judgment for 147 chemistry research groups. *Scientometrics*, *67*(3), 491–502.

Available online at www.sciencedirect.com
ScienceDirect
Journal of Informetrics 1 (2007) 47–58
Journal of INFORMETRICS An International Journal
www.elsevier.com/locate/joi

Journal self-citations—Analysing the JIF mechanism

Tove Faber Frandsen

Department of Information Studies, Royal School of Library and Information Science, Copenhagen S., Denmark

Received 29 June 2006; received in revised form 30 August 2006; accepted 5 September 2006

Abstract

This paper investigates the mechanism of the Journal Impact Factor (JIF). Although created as a journal selection tool the indicator is probably the central quantitative indicator for measuring journal quality. The focus is journal self-citations as the treatment of these in analyses and evaluations is highly disputed. The role of self-citations (both self-citing rate and self-cited rate) is investigated on a larger scale in this analysis in order to achieve statistical reliable material that can further qualify that discussion. Some of the hypotheses concerning journal self-citations are supported by the results and some are not.

Keywords: Self citations; Journal Impact Factor; Multiple linear regression

1. Introduction

The increased attention on Journal Impact Factor (JIF) as a crucial criterion of evaluation has according to Kaltenborn and Kuhn (2004) led authors and editors more or less voluntarily to adapt their publication strategy to a maximation of JIF. Editors seek to understand the impact factor calculation so that they can manipulate it to their journal's advantage (Jennings, 2001). Miller (2002), Neuberger and Counsell (2002) and Sevinc (2004) all reported that a manuscript submitted was returned by the editor requesting the author to add irrelevant references from that journal. This implies that the risk of editors manipulating JIF by increasing the number of journal self-citations is present.

The use and importance of journal self-citations is highly debated. The treatment of self-citations in analyses and evaluations has been discussed heavily and relates to how we should interpret self-citations. According to Hyland (2003) repeated self-citation accentuates one's credibility or expertise and may perpetuate one's interpretations or opinions of specific research findings or general constructs. According to Gami, Montori, Wilczynski, and Haynes (2004) critics of the impact factor (IF) as a metric of journal importance have noted the bias that results from journal self-citation but little is known about the impact of self-citations. However we cannot just ignore the existence of journal self-citations. According to Van Raan (1998b) self-citations cannot be neglected and it is necessary to perform corrections to avoid distortions. White (2001) also stressed that self-citations are not an insurmountable difficulty as they can be excluded from the analyses. But nevertheless self-citations are most often included in the calculation of JIF and could potentially have an effect on the results. Aksnes (2003) pointed out that on aggregated levels such as on national levels self-citations do not pose a problem assuming that they level out but at lower levels self-citations could potentially be a serious problem as there are great variances among, e.g. disciplines.

Models for interpreting the self-citation rates have been suggested (Rousseau, 1999) and a few investigations exist that relate self-citations to JIF (Fassoulaki, Paraskeva, Papilas, & Karabinis, 2000; Smart & Elton, 1982). But the former

E-mail address: tff@db.dk.

1751-1577/$ – see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.joi.2006.09.002

only included the self-cited rate and the latter consisted of limited data material. The main objective of this analysis is to relate self-citations to JIF. The role of self-citations (both self-citing rate and self-cited rate) will be investigated on a larger scale in this analysis in order to achieve statistical reliable material that can further qualify the discussion.

The paper is structured as follows: Section 2 surveys the research already existing within this field. Section 3 then presents and discusses the collected data and the chosen methods, followed by Section 4 with the results of the analysis. Section 5 contains conclusions and a discussion of the perspectives of the paper.

2. Overview of the existing literature

A large corpus of earlier research exists regarding self-citations and in order to keep some overview of the research we divide it into theoretically oriented and empirically oriented research. Please note that the focus here is on journal self-citations and not so much on other variations of self-citations such as author self-citations, country self-citations and institution self-citations (Eto, 2003). We begin with the theoretically oriented research noting that there are several interesting suggestions on how to evaluate and understand self-citations.

Self-citations have been translated into quantifiable measures in various forms. Rousseau (1999) defined a journal self-citation as a paper published in a journal citing papers published in the same journal. Self-citations and thus also journal self-citations have been classified by several. Lawani (1982) divided self-citations into two types called synchronous and diachronous self-citations. The synchronous rate is calculated as the citations to itself relative to the total number of references in the journal. The diachronous rate is calculated as the journal's number of self-citations relative to the total number of citations received by the journal. According to Lawani (1982, p. 282) the former is not necessarily an expression of egoism whereas on the other hand the latter is an expression of egoism as we see little or no recognition from other journals. White (2001) questioned this use of the two indicators of self-citations stating that "[w]hile Lawani's approach is intriguing, he reads egotism, usually a durable quality of personality, into data that are beyond authors' control and whose proportions can change: a sudden influx of citations from others could turn today's monster of vanity into a decent, humble fellow overnight. The charge should perhaps be reserved for failings more clearly personal, such as citing one's own work when it is irrelevant. I would even argue that abnormally high self-citation in [. . .] the synchronous rate [. . .] would be a better measure of egotism; at least it would reflect behaviour attributable to the citer. A high diachronous self-citation rate, on the other hand, seems more an indicator of what might be called intellectual isolation (true egotists prefer undiscoveredness). In any case, if egotism is defined as excessive self-citation, the burden of proving excess is on the definer". Egghe and Rousseau (1990) classified self-citations in two indicators: self-citing rate and self-cited rate. Self-citing rate relates a journal's self-citations to the total number of references it gives. Self-cited rate relates a journal's self-citations to the number of times it is cited by all journals, including itself. So basically the only thing differentiating the typologies by Lawani and Egghe and Rousseau is the terminology. Here, we choose to use the terminology suggested by Egghe and Rousseau (1990) and the mathematical definition of self-citations based on Garfield (1974) is illustrated by Table 1.

Journal J cites itself a times; it cites other journals b times. Journal J is cited by other journals c times. The self-citing rate is $a/(a+b)$; the self-cited rate is $a/(a+c)$.

Rousseau (1999) furthermore suggested that a high self-cited rate could be an expression of low visibility of the journal. Self-cited rates of leading journals would be expected to be low and the other way around for more peripheral journals. A high self-citing rate on the other hand is a sign of low visibility of the field covered by the journal and journals with high self-citing rates would tend to be more specialised. He also stressed that one should note that self-citations typically contain a different kind of information than other citations as they often contain intra-journal information (Rousseau, 1999). According to Gami et al. (2004) self-citations serve necessary functions. It allows expanding on

Table 1
Journal self-citation table

Citing journals	Cited journals	
	J	O
J	a	b
O	c	–

previous hypotheses, refer to established study designs and methods, and justify further investigations on the basis of prior results. Hence, he argues, self-citations may be inevitable when the published data are only published in a single journal.

Turning to the more empirically oriented research we note that little work exist that relate self-citations to JIF on a larger scale in order to test some of the models of interpretation suggested. Several empirical analyses focus on journal self-citations and some of these are: Pichappan (1995) indicated that the self-citing rate of a journal is affected not only by the length of existence of the journals, but also by the source articles of the journal cited and citing it. Snyder and Bonzi (1998) showed that motives to self-cite are the same as motives to cite others. Furthermore, they showed that there are large differences in the number of self-citations among disciplines but the number is constant within disciplines. Van Raan (1998a) stressed the importance of the size of the data material. It is very unlikely that all authors have the same biases and therefore a large dataset will reduce the differences. Van Raan (1998b) and Moed (2000) showed that the impact of research results is affected by the degree of international cooperation but the increased impact is not exclusively due to self-citations because even after correcting for self-citations the impact is still greater. This finding was also supported by Aksnes (2003).

Rousseau (1999) investigated the amount of self-citations in 10 highly estimated journals and 10 randomly chosen journals. He finds that self-citations are given earlier after publication than non-self-citations. Fassoulaki et al. (2000) investigated six journals for both self-citing rate and self-cited rate. Although not statistically significant they find a correlation between self-citing rate and JIF. Although restricted by the limited period of analysis and only including a single journal Peritz and Bar-Ilan (2002) found a highly increasing tendency to journal self-citations. They point to an increased significance of the journal during the period as an explanation. Rousseau and Small (2005) showed an example of a cycle of citations within the same journal issue. These journal self-citations emerged on the basis of an invisible college exchanging preprints. Finally, Tsay (2006) investigated self-citations of the most productive semiconductor journals and found that high self-citing journals are usually older, more productive and higher cited than low self-citing journals.

3. Data

The analysis in the present paper is a case study based on a number of economics journals. It is necessary to collect a rather homogeneous dataset in order to keep the number of variables at a reasonable level. Glänzel and Moed (2002, p. 178) stressed that JIF is field-specific biased and therefore one way of limiting the dataset is to use journals from only one science. A group of economics journals was selected on the basis of criteria set up by Kalaitzidakis, Mamuneas, and Stengos (2003) which ensured that the journals were scientific and belonged primarily to the social science of economics. Furthermore, the journals had to be indexed throughout the entire period in Social Science Citation Index (SSCI). A sample of 32 journals fulfilling these criteria was selected randomly and is shown in Appendix A.

Preliminary searches conducted before the start of the actual analysis showed that before the mid-1980s the number of observations in the data material is too small so the initial publication period used in the analysis is 1986 as it involves data from 1984 to 1985 when calculating the synchronous JIF. The last publication period is 2002 with corresponding citation period for the 3-year diachronic JIF of 2002–2004.

An overview of the variables is available in Appendix B and a short description follows here: the number of citations is used as a dependent variable in four different versions. To extend the indications of the analyses to more than just one JIF-calculation we calculated four different JIFs. The robustness of the results does not depend on the particular JIF chosen as we employed both synchronous and diachronous JIF. The formulas for calculating both the synchronous and the diachronous JIFs are available in Frandsen and Rousseau (2005). There are two formulations for the general case depending on whether we treat each publication year differently but as we only operate with one publication year we can use either of the formulations. We employed two 2-year synchronous JIFs which means a 2-year publication period and a 1-year citation period is used. This means that the analysis will include the citations over 1 year to publications from 2 years, e.g. citations in 1986 to articles published in 1984–1985. One was calculated as done by the ISI and one also including the document type letter in the denominator as recommended by Christensen, Ingwersen, and Wormell (1997). Furthermore, we used a 3-year diachronous JIF and a 5-year diachronous JIF. The length of the citation window must be set in accordance with the degree of obsolescence of articles within the economics literature since we want to include a large percentage of the total number of citations received. Only a few investigations of obsolescence within economics have been made. One of the few is Dorban and Vandevenne (1991) and according to their investigation

we only captured 24 percent of the citations using a citation window of 4 years but in order to perform analyses on relatively recent data we had to compromise and therefore we chose the 5-year citation period as the longest.

The time variable captures a possible development over time. By adding this variable it is possible to capture if JIF in general increases or decreases over time which could be the case if the number of included journals in the citation databases increases or decreases during the period leading to more or less possible journals to cite.

The number of self-citations is described by two related but different measures. The self-citing rate relates a journal's self-citations to the total number of references it gives. Self-cited rate relates a journal's self-citations to the number of times it is cited by all journals, including itself. In this analysis we calculate them both. The self-cited rate is calculated after using the correction technique suggested by Christensen et al. (1997). As we will explain later in Section 4 there are reasons to believe that the relation between JIF and self-cited rate is not linear per se therefore we also construct a variable describing the relationship as non-linear. That is done by computing a variable as 1 divided by the self-cited rate.

As we wish to control for other factors that might influence the results we add several variables that describe other aspects of the journals included in the analysis. The variables included here are chosen as they are expected to affect the distribution of JIF across journals. Others could have been chosen and that could potentially alter the outcome of the analyses. Future analyses will have to investigate if other factors influence the JIF and we focus on variables describing document types and geographic relations. We record the composition of document types each year. The documents are divided into seven categories namely: article, review, letter, note, editorial, book review and other. The categories consist of just the document type indicated in the category label. Only exception is the category *other* that consists of *discussion*, *item about an individual* and that sort of publications. These document types have been aggregated in this category as the dataset revealed so few of them and the use of them varies considerably over the years. Furthermore, we register the total number of publications of each journal, the share of documents with scientific content (article, review, letter and note) and the number of documents included by the ISI (article, review and note).

A variable describes the geographical location of the journal and is constructed by determining the place of publication. We are primarily interested in the few journals not originating from North America in order to describe the geographical periphery of science. This geographic location of a journal is determined by using *Ulrich's International Periodicals Directory*. When using Ulrich's for determining the geographic location it can be problematic for journals published by, e.g. Elsevier who are registered in Ulrich's as being published in The Netherlands while the reality may be different. But for this analysis we have to rely on the directory, as it can be almost impossible to establish a certain geographic location. Should all journals without certain geographic location have been discarded from the analysis it would have left us with very limited material as can be seen in Frandsen (2005). The second variable concerning geographical relations is constructed in order to record the languages of the journals and is computed as the share of documents not written in English. We could also have added a variable on the geographic location of the authors publishing in the journal but as the main focus point here is not geographic relations we restrict the variables on geographic relations to the two mentioned here.

Different estimation equations were used in order to analyse the data material. A minor analysis took place before the central main analysis. But both of these analyses consisted of variables already available through the main analysis. Furthermore, we analysed the degree of self-citing rates in order to see if the degree of self-citing could be explained by some of the other variables describing the journals. We analysed self-citing rate as the dependent variable and the estimation equation we used is as follows:

$$\begin{aligned}\text{Self-citing rate}_{i,t} = {} & \beta_0 + \beta_1(\text{total number of documents}_{i,t}) + \beta_2(\text{documents included in the ISI-JIF}_{i,t}) \\ & + \beta_3(\text{geographic location of journal}_{i,t}) + \beta_4(\text{trend}_t) + \beta_5(\text{scientific content share of total}_{i,t}) \\ & + \beta_6(\text{share of non-English language}_{i,t}) + \beta_7(\text{article}_{i,t}) + \beta_8(\text{review}_{i,t}) + \beta_9(\text{letter}_{i,t}) \\ & + \beta_{10}(\text{note}_{i,t}) + \beta_{11}(\text{editorial}_{i,t}) + \beta_{12}(\text{book review}_{i,t}) + \beta_{13}(\text{other}_{i,t}) + u_{i,t}\end{aligned}$$

where i denotes the journal whereas t the time period, β_0 the constant and $u_{i,t}$ denotes the error term.

Finally, we analysed the dataset using various forms of JIFs as the dependent variable. We wanted to be able to understand and explain the actual JIF-value of each journal. The estimation equation we used is as follows:

$$\begin{aligned}\text{JIF}_{i,t} = {} & \beta_0 + \beta_1(\text{self-cited rate}_{i,t}) + \beta_2(\text{self-citing rate}_{i,t}) + \beta_3(\text{geographic location of journal}_{i,t}) \\ & + \beta_4(\text{documents included in the ISI-JIF}_{i,t}) + \beta_5(\text{total number of documents}_{i,t})\end{aligned}$$

$$+\beta_6(\text{share of non-English language}_{i,t}) + \beta_7(\text{article}_{i,t}) + \beta_8(\text{review}_{i,t})$$
$$+\beta_9(\text{letter}_{i,t}) + \beta_{10}(\text{note}_{i,t}) + \beta_{11}(\text{editorial}_{i,t}) + \beta_{12}(\text{book review}_{i,t})$$
$$+\beta_{13}(\text{other}_{i,t}) + \beta_{14}(\text{trend}_t) + \beta_{15}(\text{scientific content share of total}_{i,t}) + u_{i,t}$$

where i denotes the journal whereas t the time period, β_0 the constant and $u_{i,t}$ denotes the error term.

We have to bear in mind that we cannot compare the coefficients from one JIF regression to another as they cannot be compared across different analyses. But it gives us an opportunity to see which variables explain the JIF statistically significant and to see if the picture depicted is the same for all JIF types analysed.

For these analyses the three Dialog Classic implementations of Arts & Humanities Citation Index (A&HCI), Science Citation Index (SCI) and Social Sciences Citation Index (SSCI) have been used. All three citation databases have been used, as citations received from journals outside the home discipline are just as relevant for this study as those from within the home discipline. In the analysis we only included citations from journals covered by ISI.

The analyses below consist of different statistical analyses of the data material. Multivariate linear regression analysis of the statistical relations between the dependent and the independent variables gives information on statistically significant relations having controlled for otherwise hidden relations with other variables. Furthermore, we are given the slope coefficients and a p-value for the linear relationship. Pearson's R^2 reveals information about the degree of correlation between the dependent and the independent variables when controlling for the effects of the other variables. The analyses have been made in Microsoft Excel and SPSS.

4. Results

Before scrutinising the linear regression analyses and interpreting the coefficients it must be emphasised that when we interpret the coefficients we say: increasing a given independent variable by, e.g. 0.3 is interpreted as leading to an increase in the dependent variable by 0.3 all other things equal. However, that is not to be understood deterministic. It is only statistical tendencies in the dataset and not predictors for the future.

Table 2 is a transcript of the output of the linear regression. First of all we notice that the R-square is not very large which means that this model is not an especially good fit. The R-square is 0.312 which means that we can explain 31 percent of the variation in the dataset. That is not impressive but the regression can still provide insight into the self-citing rates. Please note that non-significant variables are not included in the table.

First of all we can see that some document types influence the self-citing rate negatively and others positively. Journals containing many articles and notes will tend to get a higher self-citing rate. On the other hand, journals consisting of many book reviews and reviews will tend to have a lower self-citing rate. This is an expected finding as these document types (and all document types in general) contain references primarily to other document types than those two types as shown by Moed and Van Leeuwen (1995).

Table 2
Multivariate linear regression analysis

Variable	Coefficients	t-Statistic	p-Value
Intercept	0.03403	10.696	<0.01
Geographic location journal	−0.00339	−1.683	<0.1
Share of publications not in English	−0.00940	−2.664	<0.01
Article	0.00008	3.216	<0.01
Review	−0.00114	−1.818	<0.1
Letter	–	–	–
Note	0.00075	4.813	<0.01
Editorial	–	–	–
Book review	−0.00008	−2.781	<0.01
Other	–	–	–
Trend	−0.00068	−1.767	<0.1
R-squared	0.312		
Observations	288		

Dependent variable is self-citing rate.

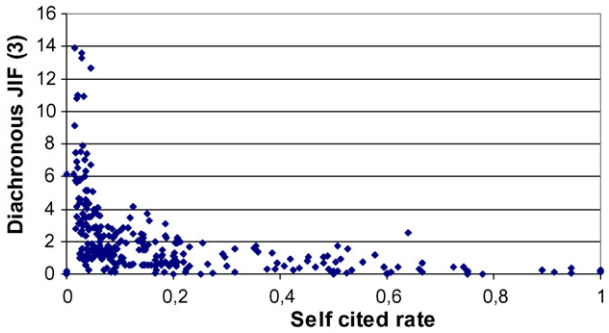


Fig. 1. Self-cited rate and the 3-year diachronous JIF.

Table 3
Multivariate linear regression analysis of JIF and independent variable is self-cited rate

JIF	*R*-square	Coefficient of dependent variable	*p*-Value of coefficient
Synchronous JIF exclusion letter	0.187	−1.994	<0.01
Synchronous JIF inclusion letter	0.188	−1.974	<0.01
Diachronous 3-year JIF	0.196	−4.274	<0.01
Diachronous 5-year JIF	0.210	−10.774	<0.01

Furthermore, we can see in the table that the geographic location of the journal influences the self-citing rates negatively. The coefficient of −0.00339 is interpreted as journals from outside North America having a self-citing rate that is 0.00339 lower than other journals. The language variable also contributes. Journals not written in English have lower self-citing rates as they have a self-citing rate that is 0.00940 lower than journals written in English. These two variables affect the self-citing rate negatively and can perhaps be explained the same way as the importance of composition of document types. There might be a tendency to citing these peripheral areas less than mainstream research which also can be detected in the self-citing rates. Journals containing many of these documents will – just as the rest of the scientific community – cite them less. But that is beyond the scope of these analyses to investigate.

A few remarks need to be made concerning self-cited rates as the relationship between JIF and self-cited rate may not be described best as linear which is default when we employ a linear regression. Fig. 1 is an illustration of the relationship between the 3-year diachronous JIF and self-cited rate. In the figure it is evident that the relationship cannot be viewed as linear.

This is just an illustration of the relationship between the 3-year diachronous JIF and self-cited rate. To extend the point to all different types of JIFs we define four different linear regression models. All with self-cited rate as independent variable and JIF as dependent. A short summary is available in Table 3.

As we can see the relationship is statistically significant for all four JIFs and in all four cases the coefficient is negative. Furthermore, it is clear that the *R*-square of all types of JIF is not very high. Therefore, we try to describe the relationship non-linearly by employing the transformed self-cited rate. The results of the new regressions can be seen in Table 4.

First of all we note the much higher *R*-square which indicates that this is a much better fit. We also note that the coefficients are no longer negative but that is due to the changing of the variable. In the further analysis we therefore choose to describe the relationship between JIF and self-cited rate as non-linear by employing the transformed version of the variable (Table 5).

Table 4
Multivariate linear regression analysis of JIF and independent variable is transformed self-cited rate

JIF	*R*-square	Coefficient of dependent variable	*p*-Value of coefficient
Synchronous JIF exclusion letter	0.452	0.102	<0.01
Synchronous JIF inclusion letter	0.456	0.101	<0.01
Diachronous 3-year JIF	0.458	0.215	<0.01
Diachronous 5-year JIF	0.510	0.566	<0.01

Table 5
Multivariate linear regression analysis of 3-year diachronous JIF

Variable	Coefficients	*t*-Statistic	*p*-Value
Intercept	0.438	1.530	0.127
Geographic location of journal	−1.198	−6.413	<0.01
Share of publications not in English	−0.816	−2.439	<0.05
Self-citing rate	19.910	3.727	<0.01
Self-cited rate (transformed)	0.187	15.305	<0.01
Scientific content (share of total)	−0.0179	−6.612	<0.01
Document types included in ISI-JIF	−0.007	−3.107	<0.01
Total number of documents	0.0171	8.472	<0.01
R-squared	0.626		
Observations	288		

Finally, we analyse the dataset using various forms of JIFs as the dependent variable as it could indicate which variables explain the JIF statistically significant and to see if the picture depicted is the same for all JIF types analysed.

In order to preserve an overview over the analyses we start out by describing the differences between the four models in order to be able to single one of the models out and describe it further. It will be time consuming and more or less purposeless to describe all four models as we show now. To illustrate the close relatedness of the four models we have constructed Fig. 2 in which we can see the rank of each journal according to JIF. As we have 32 journals times 9 time periods in the dataset we end up with ranks from 1 to 288. The circles in the figure illustrate the correlation between the two synchronous JIFs and as they are the two JIFs most similar in description they form an almost straight line. The squares in the figure compare one of the synchronous JIFs with one of the diachronous JIFs and as we can see it is not a straight line but they are closely related. The triangles in the figure illustrate the correlation between the two diachronous JIFs and again we see a close relation. This intra-disciplinary ranking with little difference between different JIFs is in accordance with results found by Garfield (1998), Moed, Van Leeuwen, and Reedijk (1999) and Stegmann (1999).

As the four models are so closely related we choose to describe only one and in cases where the models differ we describe the differences. So in the following 3-year diachronous JIF is the main focus point. First of all we start out

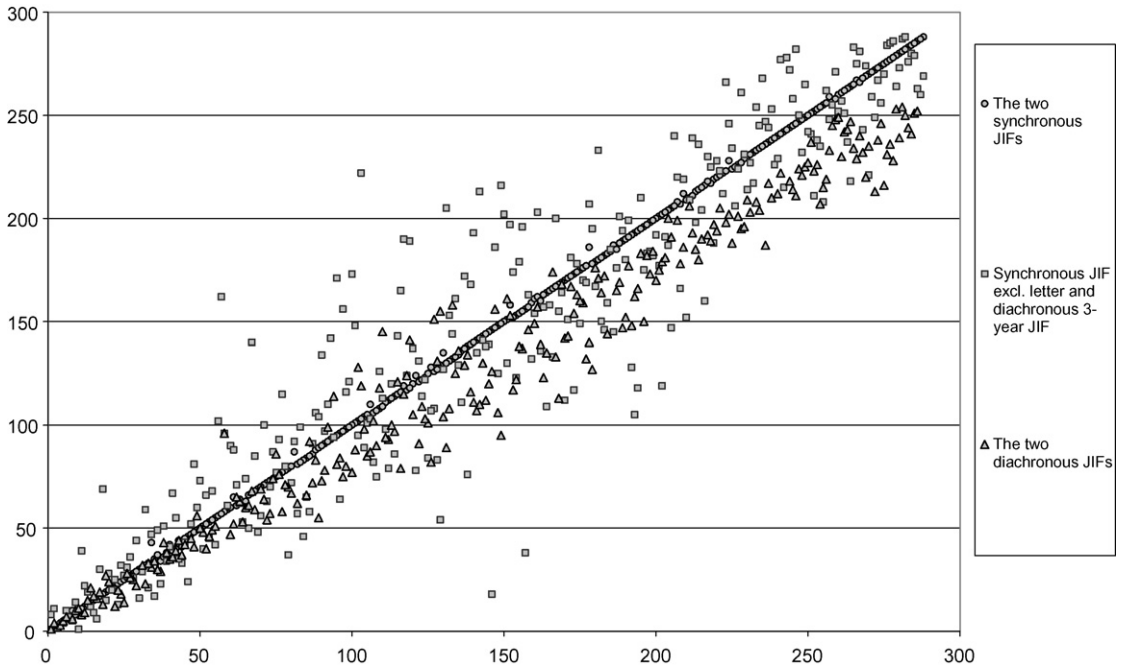


Fig. 2. Correlation between journal rankings of selected models.

by presenting the results of the linear regression model and in the following we concentrate on each element of the model.

We start by describing the influence by the time periods as described in the model. The *trend* variable is not statistically significant in the four models at the 0.1 level. That gives us an indication that we will not benefit from including in it the model. In future research and maybe including further data it might prove to be statistically significant but future research will have to cast light on that element. For the time being we can only conclude that the variable is not significant and therefore JIF is not increasing or decreasing in general over the years.

We tried running the model including all the document types but the high number of variables included weakened the model considerably and very few turned out to contribute to the understanding of JIF. Therefore, we only include the document types in aggregated forms. First of all we can see that the total number of documents significantly contributes to increasing the JIF. The coefficient of 0.0171 is to be understood like this: if a journal editor manages to increase the total number of documents published in the journal each year by 10 we will see an increase in JIF by 0.171. This finding is in accordance with Rousseau and Van Hooydonk (1996) who also found a positive correlation between the number of published articles and the impact factor.

The distribution of the total number of documents on document types show us that decreasing the share of documents containing the highest degree of scientific material will increase the JIF. The coefficient is −0.0179 and statistically significant at the 0.01 level. This means that journals with a scientific content share of 0.1 greater than another journal will have a JIF −0.00179 lower all other things equal. This aspect is further enlightened by the variable of document types included in the ISI calculation of JIF as it is also significant at the 0.01 level and the coefficient is negative. The coefficient of −0.007 means that if we increase the number included in the ISI calculation of JIF by 100 the JIF decreases by 0.7. Both variables are significant which indicates that it is not only a matter of the actual numbers of documents with scientific content it is also a matter of the share these documents comprise when we consider the total number of documents.

We notice that the variable describing the geographic origin of the journal is statistically significant at the 0.01 level. The coefficient of −1.198 should be interpreted as non-North American originated journals have JIFs that are 1.198 lower. Strong American dominance has been widely recognised as noted by, e.g. Van Dalen (1999) reporting that 44 percent of the Nobel Prize winners in economics are born outside the US, but all of these have begun their award winning work in America. There is a large export of economics researchers from the rest of the world to the United States (and Canada). Hodgson and Rothman (1999) examined the institutional background of editors and authors of 30 economic journals and also found strong American dominance. Even though the dominance is recognised this does not make the phenomenon any less interesting. Almost all economics journals describe themselves as being *international* and accept manuscripts from all over the world. As manuscripts allegedly are judged purely on their academic quality such a strong American dominance should not necessarily prevail. While North American and other journals in principle publish the same types of articles, the analysis here clearly shows that there is a difference in the degree of which these articles are cited, even when controlling for a number of factors. Such a result need to be taken into account when rankings of journals are constructed for evaluation purposes since publication in European journals will affect citation numbers downwards.

The other variable concerned with geographical relations is statistically significant and that is the variable describing the share of documents not written in English. The variable is significant at the 0.05 level and a coefficient of −0.816 tells us that increasing the share of documents not written in English will decrease the JIF. Increasing the share by 0.1 (meaning that 10 percent more of the documents are not written in English) decreases the JIF by 0.0816. However, it should be noted that according to Archambault, Vignola-Gagné, Côté, Larivière, and Gingras (2006) the SSCI selection of journals favours English and the bias affects citation analysis.

The self-citing rate variable is statistically significant at the 0.01 level and the coefficient is positive. The coefficient of 19.910 should be understood as follows: if the share of self-citations is increased by 0.1 which means that 10 percent more of the references in the journal are to the journal itself, the JIF will increase by 1.991. The self-cited rate variable is also significant at the 0.01 level. The coefficient is positive but that is due to the fact that we have transformed the original variable. This means that a positive coefficient of the transformed self-cited rate is to be understood as a negative coefficient of the self-cited rate.

As we are analysing journal self-citations as a means to a better understanding of the mechanism of JIF we will put the positive correlation between JIF and self-citing rate into perspective. We also have to be aware of the self-cited rate variable as this further complicates things. Inherent in the mathematical definitions there is a close relationship

between the two self-citation rates. Using the notation in Table 1 we can differentiate the self-citing rate with respect to a and the self-cited rate with respect to a and we find that they will both increase if a (for self-citing rate) is increased which means that an increase in the number of self-citations all other things equal will lead to increased self-citing rates and self-cited rates.

Although inherent in the mathematical definition it is a paradox that an increase in the self-citing rate leads to an increase in JIF but it also leads to an increasing self-cited rate which is related to a lower JIF. This implies that JIF cannot easily by manipulated by increasing the number of self-citations in the journal. The findings can be seen as a defence of JIF as an indicator of quality as this is how we would want the rewarding system to work. Journals acknowledged to a large extent by other journals are exponents of high quality whereas journals primarily acknowledged by themselves are not. Interpreting the results as this we can also support Lawani's theory that self-citing rate is not an expression of egoism whereas on the other hand the self-cited rate is an expression of egoism as we see little or no recognition from other journals.

However, they could also be seen as an example of the center–periphery issues in scholarly communication. It has been stated by Whitley (1991) that economics is dominated by a core of journals which maintain a particular view of economics. The periphery is engaged in alternative perceptions of economics and is not allowed to gain a foothold by the self-reinforcing hierarchy. Interpreting the results using this perspective we see a number of journals with low JIFs and high self-cited rates which we can determine as being the periphery in the set of economics journals in this analysis. They are perhaps not focused on main stream research topics and/or using a heterodox theoretical approach. The potential number of citing and cited journals is low and thus the journal is more or less isolated in the periphery of economics. On the other hand, we find a number of journals with high JIFs and low self-cited rates. They are focused on mainstream topics and/or using widely accepted theoretical approaches. Their potential number of citing and cited journals is high and thus they are a part of the dominant core maintaining the hierarchy.

How the results should be interpreted is beyond the scope of this analysis as it would require an in-depth analysis based on more qualitative investigations into structure of economics.

5. Conclusion

In this paper JIF-mechanism is investigated by focusing on journal self-citations as the treatment of these in analyses and evaluations is highly disputed. First of all we have to stress that this paper only a relatively small number of journals from only one social science. Furthermore, a number of variables are selected to be included in the study but adding more or others could potentially modify the picture depicted here.

Bearing in mind that we cannot generalise the results we can conclude that increasing the self-citing rate increases JIF. The self-citing rate is to some extent determined by the profile of the journal and has to do with the composition of document types, geographical location, language and a development over time. Furthermore, we can conclude that due to the mathematical definitions the self-citing rate and the self-cited rate are positively related. Finally, we can conclude that the transformed self-cited rate is positively correlated with JIF which is to be interpreted as an increase in the self-cited rate is related to a decrease in JIF.

Applying one perspective of analysis we can see this as a defence of JIF as is provides support to the hypothesis that JIF is capturing the impact and quality of journals. Journals acknowledged to a large extent by other journals are exponents of high quality whereas journals primarily acknowledged by themselves are not. This also gives support to Lawani's theory that self-citing rate is not an expression of egoism whereas on the other hand the self-cited rate is an expression of egoism as we see little or no recognition from other journals. However, there are other models of interpretation of the data. An alternative is to see this as a contribution to the center–periphery discussion in scholarly communication as it could be describing the characteristics of the highly cited core within economics and the low cited and isolated periphery.

Although many of the findings are as we expected we hereby provide the statistical analyses to support the hypotheses. Hopefully, this can qualify the debate on JIFs and journal self-citations.

Acknowledgments

The author wishes to thank Ronald Rousseau for his tremendous help. The author also thanks Birger Larsen and Birger Hjørland for valuable comments and suggestions.

Appendix A. Journals included in this study

No.	Journal name
1	American Economic Review
2	American Journal of Economics and Sociology
3	Brookings Papers on Economic Activity
4	Bulletin of Indonesian Economic Studies
5	Cambridge Journal of Economics
6	Desarollo Economico—Revista de Ciencas Sociales
7	Developing Economies
8	Eastern European Economics
9	Econometrica
10	Economic History Review
11	Economic Journal
12	Economica
13	Economics Letters
14	Ekonomiska Samfundets Tidskrift
15	European Economic Review
16	Explorations in Economic History
17	International Economic Review
18	Jahrbücher Für Nationalökonomie und Statistik
19	Journal of Econometrics
20	Journal of Economic Issues
21	Journal of Economic Literature
22	Journal of Economic Theory
23	Journal of political Economy
24	Kyklos
25	Oxford Economic Papers
26	RAND Journal of Economics
27	Review of Economic Studies
28	Review of Economics and Statistics
29	Scandinavian Journal of Economics
30	South African Journal of Economics
31	World Development
32	World Economy

Appendix B. Overview of variables

Variable	Values
Synchronous JIF exclusion letter	The number of citations to a journal in a given year to the publications in that journal in the previous 2 years divided by the number of articles, reviews and notes.
Synchronous JIF inclusion letter	The number of citations to a journal in a given year to the publications in that journal in the previous 2 years divided by the number of articles, reviews, letters and notes.
Diachronous JIF 3-years	The number of citations to the publications from 1 year in a journal given in 3 years divided by the number of articles, reviews, letters and notes.
Diachronous JIF 5-years	The number of citations to the publications from 1 year in a journal given in 5 years divided by the number of articles, reviews, letters and notes.
Time period	1 = 1986; 2 = 1988; 3 = 1990; 4 = 1992; 5 = 1994; 6 = 1996; 7 = 1998; 8 = 2000; 9 = 2002.
Self-citing rate	The total number of references to a journal by itself in a given year divided by the number total number of references in the journal that year.
Self-cited rate	The total number of citations to a journal in a given year given by the journal itself divided by the total number of citations to the journal in that year. Both numerator and denominator are corrected when computed.
Transformed self-cited rate	1 divided by self-cited rate.
Article	The number of articles published by a journal in a given year.

Review	The number of reviews published by a journal in a given year.
Letter	The number of letters published by a journal in a given year.
Note	The number of notes published by a journal in a given year.
Book review	The number of book reviews published by a journal in a given year.
Editorial	The number of editorials published by a journal in a given year.
Others	The number of other document types published by a journal in a given year.
Total	The total number of publications published by a journal in a given year.
Share of document types with scientific content	Number of reviews, notes, letters and articles divided by the total number of documents.
Number of publications included in the ISI calculation of JIF	Number of reviews, notes and articles.
Geographic location of journal	0 = North America; 1 = other countries.
Number of non-English language publications	The number of publications written in a non-English language in a given year in a journal.
Share of non-English language publications	The number of documents written in a non-English language divided with the total number of publications.

References

Aksnes, D. W. (2003). A macro study of self-citations. *Scientometrics*, *56*(2), 235–246.

Archambault, É., Vignola-Gagné, É., Côté, G., Larivière, V., & Gingras, Y. (2006). Benchmarking scientific output in the social sciences and humanities: The limits of existing databases. *Scientometrics*, *68*(3), 329–342.

Christensen, F. H., Ingwersen, P., & Wormell, I. (1997). Online determination of the journal impact factor and its international properties. *Scientometrics*, *40*(3), 529–540.

Dorban, M., & Vandevenne, A. F. (1991). Bibliometric analysis of bibliographic behaviours in economic sciences. *Scientometrics*, *25*(1), 149–165.

Egghe, L., & Rousseau, R. (1990). *Introduction to informetrics. Quantitative methods in library, documentation and information science*. Amsterdam: Elsevier Science Publishers.

Eto, H. (2003). Interdisciplinary information input and output of a nano-technology project. *Scientometrics*, *58*(1), 5–33.

Fassoulaki, A., Paraskeva, A., Papilas, K., & Karabinis, G. (2000). Self-citations in six anaesthesia journals and their significance in determining the impact factor. *British Journal of Anaesthesia*, *84*(2), 266–269.

Frandsen, T. F., & Rousseau, R. (2005). Article impact calculated over arbitrary periods. *Journal of the American Society for Information Science and Technology*, *56*(1), 58–62.

Frandsen, T. F. (2005). Journal interaction: A bibliometric analysis of economics journals. *Journal of Documentation*, *61*(3).

Gami, A. S., Montori, V. M., Wilczynski, N. L., & Haynes, R. B. (2004). Author self-citation in the diabetes literature. *Canadian Medical Association Journal*, *170*(13), 1925–1927.

Garfield, E. (1974). Journal citation studies. XVII. *Journal Self-Citation Rates—There's a Difference, Current Contents*, *52*(December), 1974.

Garfield, E. (1998). Long-term vs. short-term journal impacts: Does it matter? *The Scientist*, *12*(3), 10–12.

Glänzel, W., & Moed, H. F. (2002). Journal impact measures in bibliometric research. *Scientometrics*, *53*, 171–193.

Hodgson, G., & Rothman, H. (1999). The editors and authors of economics journals: A case of institutional oligopoly? *The Economics Journal*, *109*, 165–186.

Hyland, K. (2003). Self-citation and self-reference: Credibility and promotion in academic publication. *Journal of the American Society for Information Science and Technology*, *54*, 251–259.

Jennings, C. (2001). Citation data: The wrong impact? *Cortex Forum*, *37*(4), 585–589.

Kalaitzidakis, P., Mamuneas, T. P., & Stengos, T. (2003). Rankings of academic journals and institutions in economics. *Journal of the European Economic Association*, *1*(6), 1346–1366.

Kaltenborn, K. F., & Kuhn, K. (2004). The journal impact factor as a parameter for the evaluation of researchers and research. *Revista Espanola de Enfermedades Digestivas*, *96*(7), 460–476.

Lawani, S. M. (1982). On the heterogeneity and classification of author self-citations. *Journal of the American Society for Information Science*, *33*, 281–284.

Miller, J. B. (2002). Impact factors and publishing research. *Scientist*, *16*, 11.

Moed, H. F. (2000). Bibliometric indicators reflect publication and management strategies. *Scientometrics*, *47*(2), 323–346.

Moed, H. F., & Van Leeuwen, T. N. (1995). Improving the accuracy of Institute for Scientific Information's journal impact factors. *Journal of the American Society for Information Science*, *46*(6), 461–467.

Moed, H. F., Van Leeuwen, T. N., & Reedijk, J. (1999). Towards appropriate indicators of journal impact. *Scientometrics*, *46*, 575–589.

Neuberger, J., & Counsell, C. (2002). Impact factors: Uses and abuses. *European Journal of Gastroenterology and Hepatology*, *14*(3), 209–211.

Peritz, B. C., & Bar-Ilan, J. (2002). The sources used by bibliometrics–scientometrics as reflected in references. *Scientometrics*, *54*(2), 269–284.

Pichappan, P. (1995). A dual refinement of journal self-citation measures. *Scientometrics*, *33*(1), 13–21.

Rousseau, R. (1999). Temporal differences in self-citation rates of scientific journals. *Scientometrics*, *44*(3), 521–531.

Rousseau, R., & Small, H. (2005). Escher staircases dwarfed. *ISSI Newsletter*, *1*(4), 8–10.

Rousseau, R., & Van Hooydonk, G. (1996). Journal production and journal impact factors. *Journal of the American Society for Information Science*, *47*(10), 775–780.

Sevinc, A. (2004). Manipulating impact factor. An unethical issue or an editor's choice? *Swiss Medical Weekly*, *134*, 410.

Smart, J. C., & Elton, C. F. (1982). Consumption factor scores of psychology journals: Scientometric proporties and qualitative implications. *Scientometrics*, *4*(5), 349–360.

Snyder, H., & Bonzi, S. (1998). Patterns of self-citation across disciplines (1980–1989). *Journal of Documentation*, *55*(5), 431–435.

Stegmann, J. (1999). Building a list of journals with constructed impact factors. *Journal of Documentation*, *55*(3), 310–324.

Tsay, M.-Y. (2006). Journal self-citation study for semiconductor literature: Synchronous and diachronous approach. *Information Processing and Management*, *42*(6), 1567–1577.

Van Dalen, H. (1999). The golden age of nobel economists. *The American Economist*, *43*(2), s.19–s.35.

Van Raan, A. F. J. (1998a). In matters of quantitative studies of science the fault of theorists is offering too little and asking too much. *Scientometrics*, *43*(1), 129–139.

Van Raan, A. F. J. (1998b). The impact of international collaboration on the impact of research results. *Scientometrics*, *42*(3), 423–428.

White, H. (2001). Authors as citers over time. *Journal of the American Society for Information Science and Technology*, *52*(2), 87–108.

Whitley, R. (1991). *The organisation and role of journals in economics and other scientific fields*. Working Paper 204. Manchester business School.

Available online at www.sciencedirect.com
ScienceDirect
Journal of Informetrics 1 (2007) 68–82
Journal of INFORMETRICS
An International Journal
www.elsevier.com/locate/joi

Mapping the bid behavior of conference referees

Marko A. Rodriguez[a,*], Johan Bollen[b], Herbert Van de Sompel[b]

[a] *CCS-3: Knowledge & Information Systems Science Team, Los Alamos National Laboratory, United States*
[b] *Digital Library Research & Prototyping Team, Los Alamos National Laboratory, United States*

Received 7 July 2006; received in revised form 14 September 2006; accepted 19 September 2006

Abstract

The peer-review process, in its present form, has been repeatedly criticized. Of the many critiques ranging from publication delays to referee bias, this paper will focus specifically on the issue of how submitted manuscripts are distributed to qualified referees. Unqualified referees, without the proper knowledge of a manuscript's domain, may reject a perfectly valid study or potentially more damaging, unknowingly accept a faulty or fraudulent result. In this paper, referee competence is analyzed with respect to referee bid data collected from the 2005 Joint Conference on Digital Libraries (JCDL). The analysis of the referee bid behavior provides a validation of the intuition that referees are bidding on conference submissions with regards to the subject domain of the submission. Unfortunately, this relationship is not strong and therefore suggests that there may potentially exists other factors beyond subject domain that may be influencing referees to bid for particular submissions.

Keywords: Peer-review; Referees; Bid behavior

1. Introduction

The peer-review process is the most widely accepted method for validating research results within the scientific community. However, its credibility as a valid certification mechanism has come under scrutiny. There exists a rich body of literature that points to many of the inadequacies of the current system (Evans, 1995; El-Munchid, 2001; Bence & Oppenheim, 2004), but of particular interest to this paper is the issue concerned with ensuring that referees are in fact reviewing manuscripts within their domain of expertise (Kassirer & Campion, 1994; Eisenhart, 2002). There exists a series of stages within the peer-review process that ultimately lead up to a referee review. One of the first and potentially most important stage is the one that attempts to distribute submitted manuscripts to competent referees. Unfortunately, it is difficult to study many of the stages of the peer-review process due to its confidential nature. Therefore, much of the peer-review process, including referee assignment, remains sheltered from the rigors of the scientific method. Fortunately, the program chairs and steering committee of the 2005 Joint Conference on Digital Libraries[1] (JCDL) has provided the Los Alamos National Laboratory (LANL) Digital Library Research and Prototyping team the referee bid data used for their 2005 conference peer-review process so that referee assignment could be analyzed for this study.

* Corresponding author. Fax: +1 505 665 6452.
E-mail addresses: marko@lanl.gov (M.A. Rodriguez), jbollen@lanl.gov (J. Bollen), herbertv@lanl.gov (H. Van de Sompel).
[1] JCDL 2005 is located at: http://www.jcdl2005.org/.

1751-1577/$ – see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.joi.2006.09.006

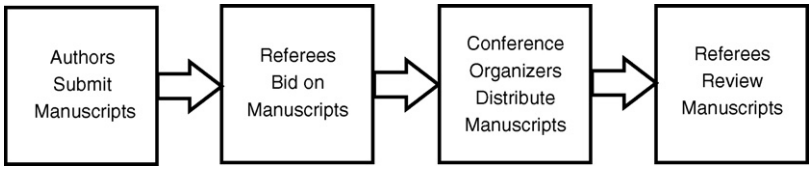


Fig. 1. Typical conference review stages.

In conference situations, where there exist a large number of submissions at one particular point in time (near the submission deadline date), conference organizers tend to rely on a pool of pre-selected referees to review the submission archive. The conference organizers require each referee to briefly look over each submission (e.g. read each submission abstract or ACM classification codes) and place submission bids. A referee bid states the referee's subjective opinion of their level of expertise with regards to a submission. Furthermore, conflict of interest situations are usually identified at this point. Once all the referee bids have been collected, the conference organizers can use any number of the many documented manuscript-to-referee matching algorithms to distribute each submission to a set of competent referees (Wei, Hartvigsen, & Czuchlewski, 1999). These stages are represented in Fig. 1. The data set provided by the 2005 JCDL program chair does not state which referees reviewed which submission, only the subjective opinion of the referee's level of expertise with respect to each submission.

Since conference organizers ask their referees to bid on submissions with regard to their domain of expertise, it is hypothesized that referee bidding is based on two factors: (1) the subject domain of the submission and (2) the expertise of the referee. The validity of this hypothesis is investigated using various statistical techniques that rely on a keyword analysis of submission abstracts and the location of each referee within the greater scientific community's co-authorship network. In short, the analysis demonstrates that the referees of the 2005 JCDL program committee are, in fact, bidding for submissions with respect to the subject domain of the submissions. Unfortunately, the strength of this relationship is not strong enough to conclude that submission subject domain is the only, or even the most significant, factor influencing referee bidding behavior.

2. The 2005 JCDL bid data set

The JCDL is an international forum that focuses on the technical, practical, and social issues concerning digital libraries. Each year, the JCDL hosts a conference to present technical papers, posters, demonstrations, tutorials, etc., that present recent developments in the digital library community. From June 7 to June 10 of 2005, the JCDL was held in Denver, Colorado in the United States (Sumner, 2005). The bid data provided by the 2005 JCDL program chair is considered extremely sensitive, therefore careful handling and analysis of this data was the first priority of this research endeavor. All information that is not publicly available from the JCDL website is, to the best of our knowledge, indeterminable from the presented results. Information, such as which submissions were rejected is not provided. The names of the referees have been anonymized by assigning each referee a unique random identifier. This section will discuss the bid data provided by the 2005 JCDL program chair as well as the various manipulations necessary to appropriately represent this information for analysis.

There were 264 submissions to the 2005 JCDL. Of those 264 submissions, 105 were full technical articles, 77 were short technical articles, 40 were posters, 17 were demonstrations, 4 were panel talks, 7 were tutorials, 7 were workshop talks, and 7 were doctoral presentations. The JCDL program committee provided the authors a table containing each submission's unique identification number, title, authors, type, and acceptance/rejection status. An example subset of this data is provided in Table 1. The submission titles and authors of those submission that were rejected by the committee have been replaced with the ### notation in order to protect the privacy of the submitters. Since accepted submissions are freely accessible, information pertaining to accepted publications is provided[2]. Furthermore, note that the title and authors have been truncated to ensure that the table fits within the margins of this paper.

Each referee on the 2005 JCDL program committee was asked to bid on which submissions they wished to review in terms of their expertise in the subject domain of the submission. Therefore, accompanying the submission data table

[2] JCDL 2005 proceedings located at: http://www.informatik.uni-trier.de/ ley/db/conf/jcdl/jcdl2005.html.

Table 1
Sample of the 2005 JCDL submission data

Sub id	Submission title	Submission authors	Submission type	Submission status
13	###	###	Full Technical Article	Rejected
14	###	###	Full Technical Article	Rejected
15	Creating an Infrastructure for Collaboration...	R. David Lankes,...	Short Technical Article	Accepted
16	Graph-based Text Representation Model...	Hidekazu Nakawatase,...	Full Techinical Article	Accepted
17	An Evaluation of Automatic Ontologies...	Aaron Krowne,...	Full Techinical Article	Accepted

Table 2
Example bid matrix for each submission for each program committee referee, **B**

Sub/ref	1	2	3	4	5
13	1	2	2	3	3
14	2	3	2	3	3
15	4	2	3	1	1
16	3	3	1	2	0
17	1	3	2	3	3

Table 3
The meaning of the bid values within the bid matrix, **B**

Bid	Meaning of the bid value
0	Did not provide a bid
1	Expert in the domain of the submission and wants to review
2	Expert in the domain of the submission
3	Not an expert in the domain of the submission
4	Conflict of interest between referee and submission

Table 4
The meaning of the bid values within the modified bid matrix, $\mathbf{B}'$

Bid	Meaning of the bid value
0	Unknown expertise (wildcard)
1	Expert in the domain of the submission
2	Not an expert in the domain of the submission

there also exists an associated bid matrix, $\mathbf{B} \in \mathbb{B}^{|S|\times|R|}$, where S is the set of submissions, R is the set of referees, and $\mathbb{B} = \{0, 1, 2, 3, 4\}$. It is important to note that $|S| \gg |R|$. The rows of the bid matrix refer to the unique id of each of the submissions. The columns of the bid matrix refer to the referees of the program committee. The matrix entries are the bid values provided by each referee for each submission. Therefore, $b_{i,j}$ refers to referee j's bid for submission i, where $0 \leq b_{i,j} \leq 4$. Table 2 is an artificial example of the supplied bid information. Note that the bid values for Table 2 were randomly generated and the referee names are not provided. The actual program committee for the JCDL is public information[3], but their respective bid vectors are not.

The values of the bid matrix, **B**, are not on an interval scale, but instead are nominal (i.e. each value symbolizes a particular bid type). Table 3 provides the meaning for each of the bid values.

The bid matrix provided by the 2005 JCDL, **B**, contains extraneous information, such as 'wants to review' ($b = 1$) and 'conflict of interest' ($b = 4$). Since this study focuses specifically on referee expertise, this information will be discarded. Therefore, bid categories 1 and 2 will be considered the same and bid categories 0 and 4 will be considered wildcards. The modified bid matrix used throughout the remainder of this study has the properties of $\mathbf{B}' \in \mathbb{B}'^{|S|\times|R|}$ where $\mathbb{B}' = \{0, 1, 2\}$. Table 4 has the bid meanings of the modified bid matrix.

The original artificial bid matrix provided in Table 2 is thus transformed into the one shown in Table 5.

Of the 264 submissions, only 118 of the submissions have actual bid data. This means that 146 submissions had bids of all 0. Therefore, only those submissions with a complete set of bid data will be analyzed for the remainder of

[3] JCDL 2005 program committee available at: http://www.jcdl2005.org/progcomm.html.

Table 5
Example modified bid matrix for each submission for each program committee referee, $\mathbf{B}'$

Sub/ref	1	2	3	4	5
13	1	1	1	2	2
14	1	2	1	2	2
15	0	1	2	1	1
16	2	2	1	1	0
17	1	2	1	2	2

this study. In addition, of the 76 program committee members of the JCDL, 11 members gave no bid information. No bid information is defined as an individual whose bid vector is all 0s. These referees were removed from the analysis. Finally, since a portion of this analysis is based on co-authorship behavior, those referee committee members not located within the DBLP[4] were not included in this study. Of the remaining 65 referees, 5 were not in the DBLP. Therefore, the bid matrix as defined for the remainder of this study has 118 rows (submissions), and 60 columns (referees), $\mathbf{B}' \in \mathbb{B}'^{118\times 60}$.

3. The methodology

Intuitively, when ignoring conflict of interest situations, referee bidding should be based on two factors: (1) the domain of the submission and (2) the domain of expertise of the referee. Therefore, the referee bid matrix should be the result of each referees analysis of the submission abstracts and the referee's area of expertise (their location in the scientific community's co-authorship network). This idea, which is the hypothesis of this study, is represented by the arced dotted lines at the top of Fig. 2. To verify or falsify this hypothesis, a collection of statistical techniques are used to determine the relationship between referee bidding and submission subject domain. The two factors of the hypothesis are explored according to Tracks 1 and 2 of Fig. 2.

Track 1 provides a correlation between two submission similarity matrices. The first similarity matrix is constructed using referee bid data, $\mathbf{S}_b$, and the second is constructed according to an submission abstract term analysis, $\mathbf{S}_t$ (Section 4). If referees are in fact bidding according to the subject domain of the submissions, then the correlation between $\mathbf{S}_b$ and $\mathbf{S}_t$ should be high. If the correlation is negative, or extremely low, then other factors that may not include submission subject domain are influencing referee bidding. Furthermore, it is possible to cluster submissions according to referee bid behavior. A intra-term analysis of these clusters provide an entropy value for each of the clusters. If the clusters created by referee bidding maintain a low entropy for their highest weighted terms and a low correlation between their term vectors, then it can be argued that referee bidding is driven by submission subject domain.

Track 2 provides the correlation between a referee similarity matrix created according to referee bidding behavior, $\mathbf{R}_b$, and a referee similarity matrix created using a relative-rank algorithm within a co-authorship network, $\mathbf{R}_g$ (Section 5). A high correlation means that referees who are similar in expertise, as determined by their place in the co-authorship network, are also bidding similarly. A high correlation would be expected if referee bidding is based solely on submission subject domain. If this correlation is low, then other factors besides submission subject domain are influencing referee bidding. This paper will first explore Track 1 and then Track 2.

4. The bid matrix and submission similarity

This section will present the Track 1 analysis represented in Fig. 2. In order to determine the relationship between referee bidding and submission subject domain, the submissions are related according to the bid behavior of the program committee referees, $\mathbf{S}_b$, and are related according to their abstract term-frequency inverse document-frequency (TFIDF) term weight distributions, $\mathbf{S}_t$ (Salton, 1998). In short, a TFIDF calculation determines the most descriptive words within a document (or document cluster) with respect to the entire document corpus. This section will first discuss the construction of $\mathbf{S}_b$ and then $\mathbf{S}_t$.

Since the values of the bid matrix, $\mathbf{B}'$, refer to semantic categories and not a gradient scale, a Hamming distance function is used to determine the similarity of any two submissions (Hamming, 1950). Hamming distance is defined

[4] Digital Bibliography and Library Project available at: http://www.informatik.uni-trier.de/ ley/db/.

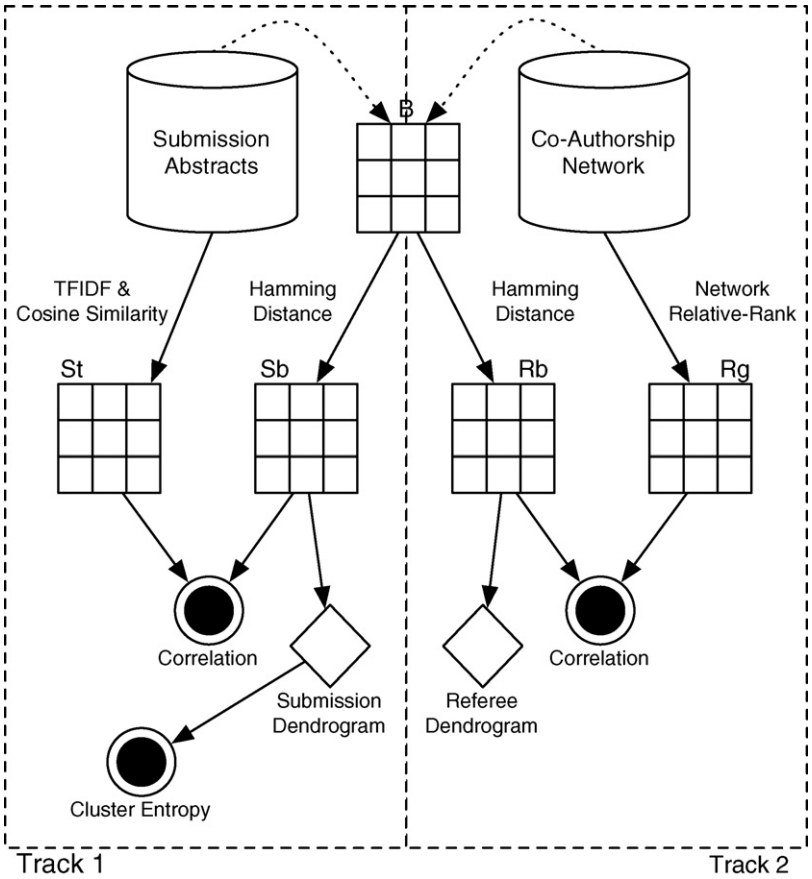


Fig. 2. Experiment outline.

as the amount of characters that differ between two strings of equal length. For example, if there exists the strings "2212" and "1212", the Hamming distance is 1 since only their first characters differ. Given the Hamming distance between two bid vectors, $h(\vec{b'}_i, \vec{b'}_j)$, and the length of a vector, $l = |\vec{b'}_l|$, the similarity between any two submissions is calculated according to Eq. (1). To account for wildcard bids ($b'_{i,j} = 0$), if any one of the two bid vectors being compared has an entry that contains a 0, that particular entry on both vectors is ignored and both their vector lengths, l, are reduced by 1. For example, when comparing the two bid vectors "0121" and "2120", their length, l, is 2 and their Hamming distance, h, is 0 because both their first and last entries are ignored and their second and third entries are equal. Therefore, their similarity is 1.

$$\mathbf{S}_{\mathrm{b}i,j} = \mathbf{S}_{\mathrm{b}j,i} = 1 - \frac{h(\vec{b'}_i, \vec{b'}_j)}{l} \tag{1}$$

Eq. (1) ensures a symmetrical submission similarity matrix, $\mathbf{S}_\mathrm{b} \in \mathbb{R}^{|S|\times|S|}$, whose diagonal values are 1.0. According to the sample bid matrix presented in Table 5, the submission similarity matrix shown in Table 6 is constructed using Eq. (1).

Table 6
Submission similarity determined according to their Hamming distance

Id	13	14	15	16	17
13	1.0	0.8	0.25	0.25	0.8
14	0.8	1.0	0.0	0.5	1.0
15	0.25	0.0	1.0	0.66	0.0
16	0.25	0.5	0.66	1.0	0.5
17	0.8	1.0	0.0	0.5	1.0

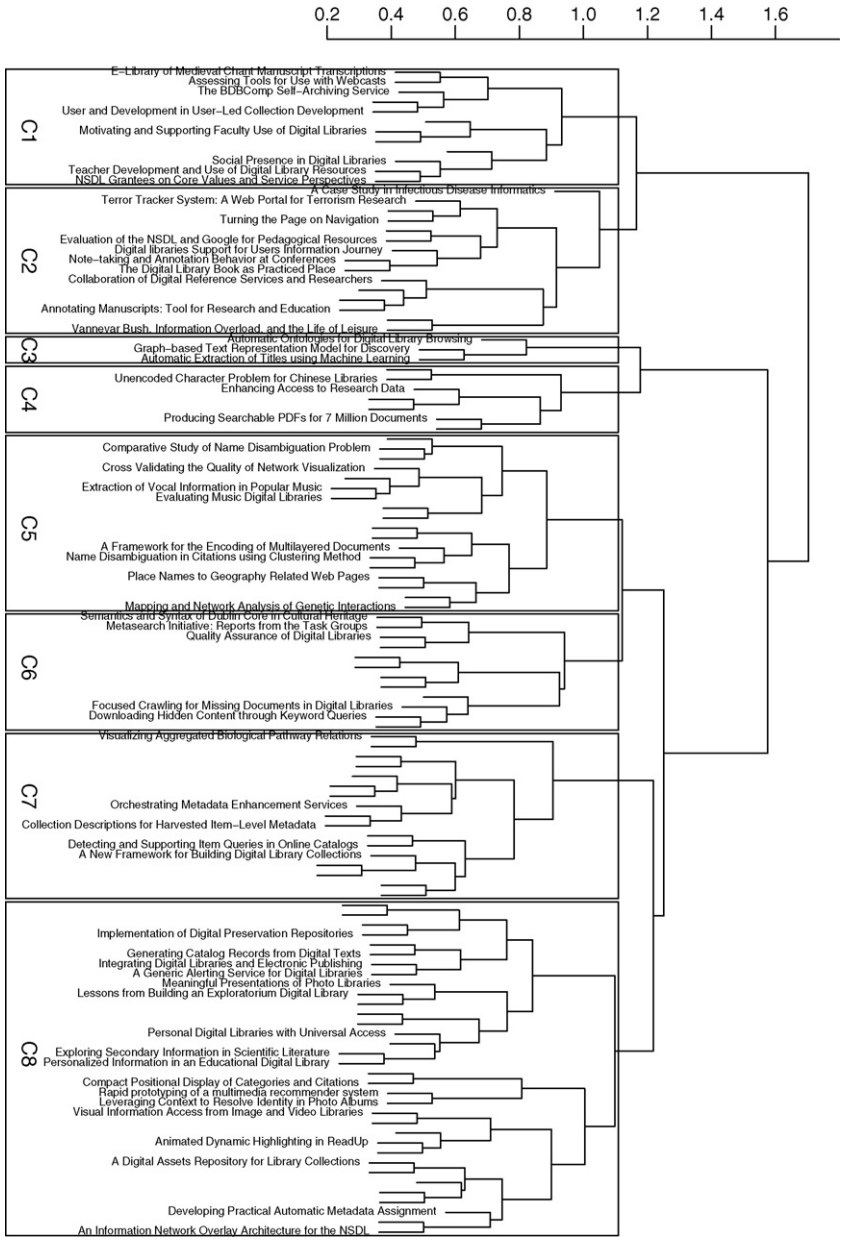


Fig. 3. Submission similarity represented according to a hierarchical cluster.

4.1. Submission similarity and the dendrogram

Once a submission similarity matrix, $\mathbf{S}_b$, has been constructed it is possible to hierarchically structure the submissions into a dendrogram in order to visualize the relationship between the various submissions by means of the complete linkage technique (Lance & Williams, 1967). The submission dendrogram constructed from $\mathbf{S}_b$ is presented in Fig. 3. Note that the titles of the rejected submissions have been left out. Accepted submission titles have been truncated to ensure readability. Furthermore, larger cluster patterns are represented as the eight boxed sections and are denoted C1 through C8. These clusters were extracted from the dendrogram by setting a threshold on the dendrogram tree height. The threshold, which is 1.1, was arbitrarily selected to expose enough clusters to make the following analysis interesting.

Table 7
Cluster feature vectors of the keywords in the submission abstracts

Cluster/term	Browser	Built	Bureau	Bush
3	3	7	3	1
4	4	3	2	0
5	1	0	1	0

A manual review of the clusters with respect to the submissions they contain demonstrates a congruency between submission topic and referee bidding. To validate this qualitative claim, three statistical techniques are used. The first involves analyzing the abstracts of the submissions of each of the clusters in order to determine cluster subject domain. The second involves determining the entropy value of each cluster. Clusters that are more strict with respect to a particular subject domain will tend to have a lower entropy. The third technique provides a correlation between a similarity matrix constructed from the cosine similarity of the TFIDF term weight vectors of each submission, $\mathbf{S}_t$, and the matrix constructed from the referee bid behavior, $\mathbf{S}_b$. This final correlation provides a single quantitative value expressing the relationship between submission subject domain and referee bidding.

4.2. Entropy in the submission clusters

The eight major clusters of the submission dendrogram derived according to referee bid data can be validated as meaningful categorizations by analyzing the terms of the submission abstracts. This requires that all submission abstracts be parsed to determine the full collection of keywords across all submission abstracts. Each abstract is processed by removing stop words and then applying the Porter stemming algorithm (Porter, 1980). These two processes remove overly frequent words (i.e. the, and, it) and perform suffix stripping (i.e. computer and computation stem to comput), respectively. Each time a particular term in the full collection of keywords is used in an abstract of one of the cluster submissions, the term frequency for that term in that cluster is incremented by 1. An example feature vector is provided in Table 7. For example, for all the submissions in cluster 3, the term *built* was used seven times.

For each cluster i, it is possible to determine how specific a particular term j is to that cluster according to Eq. (2) where $\text{freq}(i, j)$ is the frequency of term j in cluster i, $n(i)$ the total number of terms in cluster i, N the number of clusters (which is always 8 for this experiment), and $n_c(j)$ is the number of clusters for which term j appears (Salton, 1998).

$$\text{TFIDF}(i, j) = \frac{\text{freq}(i, j)}{n(i)} \times \log_{10}\left(\frac{N}{n_c(j)}\right) \tag{2}$$

The higher the TFIDF weight for term j in cluster i, the more specific term j is to the cluster i and therefore the more suited it is as a description of the cluster's subject domain. The following table presents the TFIDF calculations for the sample feature vector presented in Table 8.

In order to determine the subject domain of each of the eight clusters, the top 10 TFIDF weighted terms were extracted. Table 9 provides these terms ordered by their TFIDF weight where term 1 has a higher weight than term 2. The term weight distributions derived from the TFIDF calculation of the cluster abstracts can now be represented according to their internal cluster information content. Internal cluster information content can be calculated using the standard entropy equation as defined according to its information theoretic sense (Shannon, 1948). The lower the entropy, the more specialized, or focused, the cluster. The higher the entropy, the less specialized. Since clusters vary in size, the entropy for a cluster is calculated only for the top 10 term weights presented in Table 9. Furthermore, since an entropy calculation is defined for a probability distribution, Eq. (3) normalizes the top 10 term weights.

Table 8
Cluster TFIDF term weight vectors of the keywords in the submission abstracts

Cluster/term	Browser	Built	Bureau	Bush
3	0.00	0.08	0.00	0.03
4	0.00	0.05	0.00	0.00
5	0.00	0.00	0.00	0.00

Table 9
Top 10 terms for the eight clusters defined in Fig. 3

	C1	C2	C3	C4	C5	C6	C7	C8
1	webcast	behavior	extract	patent	name	hidden	ecl	photo
2	cyberinfrastructur	drew	powerpoint	tobacco	surrog	crawler	morf	preserv
3	ncknow	note	handel	invent	music	subschema	relev	video
4	interview	overload	mainli	determin	disambigu	flora	item-level	european
5	teacher	engag	train	hidden	segment	ontos	network	alert
6	descriptor	factor	graph	control	candid	queri	citat	dark
7	faculti	gather	step	american	network	expans	circleview	region
8	lesson	school	weight	chemic	tempor	homepag	dlii	busi
9	survei	teamsearch	algebra	compani	citat	plant	extract	addit
10	transcript	visualis	basi	searchabl	genet	reusabl	meta-inform	mobil

$$\mathrm{TFIDF}'(i, j) = \frac{\mathrm{TFIDF}(i, j)}{\sum_{k=0}^{k<10} \mathrm{TFIDF}(i, k)} \tag{3}$$

The entropy of a cluster is then calculated over the probability distribution as described by Eq. (4), where $H(i)$ is the entropy for cluster i.

$$H(i) = -\sum_{j=0}^{j<10} \mathrm{TFIDF}'(i, j) \log_2(\mathrm{TFIDF}'(i, j)) \tag{4}$$

The entropy values for the eight clusters of the dendrogram presented in Fig. 3 are presented in Table 10.

It is interesting to note that C5, the lowest entropy cluster, is composed mainly of submissions associated with name disambiguation and music in digital library research. On the other hand, the highest entropy cluster C2 has a mix of more unrelated submissions ranging from digital libraries in educational settings to infectious diseases and terrorism. Figs. 4 and 5 present the distribution of the term weights for the top 10 terms of the eight clusters. The steeper the distribution tail, the lower the cluster entropy and therefore the more focused the cluster is towards its higher weighted terms. The analysis of the terms for each cluster points to a qualitative relationship between referee bidding and submission subject domain.

A more quantitative validation can be determined when the TFIDF term weight vectors of the eight clusters are compared using a Spearman rank-order correlation. The correlations are performed on the cluster's TFIDF term weight vectors which contain the entire abstract dictionary, D, where $|D| = 2121$. Table 11 provides the Spearman rank-order correlations for each cluster comparison. What is noticeable from Table 11 is that all the correlations are less than ± 0.12. The fact that the clusters, which are organized by referee bidding, yield very low correlations between their TFIDF term weight vectors means that the clusters are well separated according to their term distributions. If these correlations were high, then it would be difficult to claim that the clusters are organized according to subject domain and thus referee bidding would not be related to submission subject domain. Since the correlations are all less that ± 0.12, this confirms the hypothesis that there does exist a relationship between the bidding behavior of the conference referees and the subject domain of the submission abstracts. The next section will further explore the strength of this relationship.

Table 10
Entropy values for the eight clusters defined in Fig. 3

Cluster	Entropy
1	3.2668
2	3.2840
3	3.2148
4	3.2213
5	3.2025
6	3.2281
7	3.2610
8	3.2442

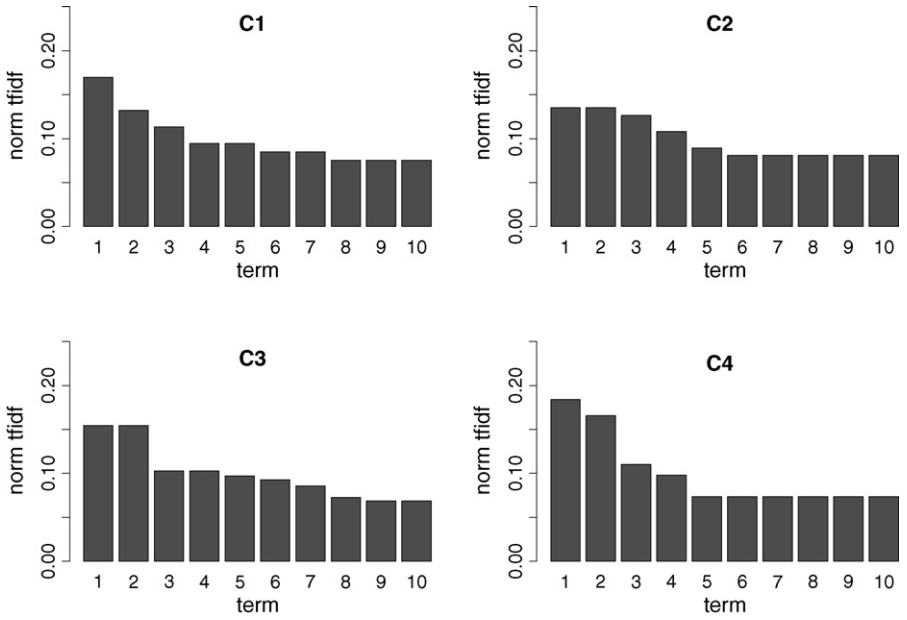


Fig. 4. Normalized TFIDF weights for the 10 terms of Table 9 for the clusters 1 through 4 defined in Fig. 3.

4.3. Cosine similarity correlation

To further quantify the relationship between referee bidding and a submission's subject domain, it is possible to correlate the relationship between submissions based on referee bidding, on the one hand, and the relationship between submissions based on their TFIDF term weight vectors, on the other. This requires the construction of the similarity matrix $\mathbf{S}_t$, which denotes the cosine similarity between every submission with respect to their complete TFIDF term weight vector. This means that each term in the abstracts of each submission is analyzed according to the TFIDF

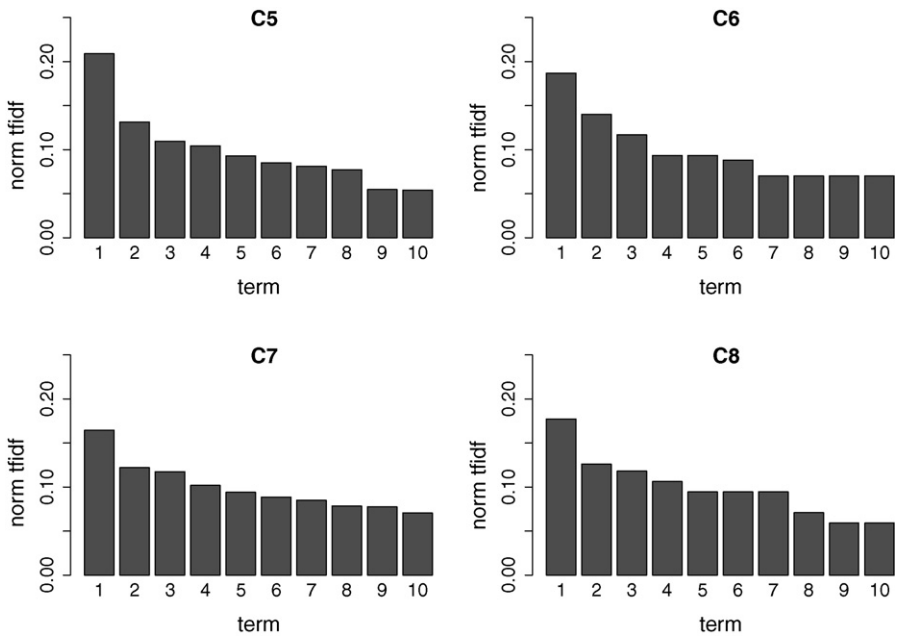


Fig. 5. Normalized TFIDF weights for the 10 terms of Table 9 for the clusters 5 through 8 defined in Fig. 3.

Table 11
Spearman rank-order correlations for the 2121 TFIDF term weights of the eight clusters defined in Fig. 3

	C1	C2	C3	C4	C5	C6	C7	C8
C1	1.0	0.10008	0.04776	0.01666	0.02687	0.08817	0.07590	0.00653
C2	0.10008	1.0	0.07817	0.03279	0.03958	0.07983	0.08225	−0.06390
C3	0.04776	0.07817	1.0	0.04314	0.10796	0.11926	0.14579	0.02117
C4	0.01666	0.03279	0.04314	1.0	0.04393	0.07461	0.06287	−0.04580
C5	0.02687	0.03958	0.10796	0.04393	1.0	0.09844	0.11082	−0.05910
C6	0.08817	0.07983	0.11926	0.07461	0.09844	1.0	0.14143	−0.00126
C7	0.07590	0.08225	0.14579	0.06287	0.11082	0.14143	1.0	0.00549
C8	0.00653	−0.06390	0.02117	−0.04580	−0.05910	−0.00126	0.00549	1.0

equation presented in Eq. (2). This results in a matrix $\mathbf{T} \in \mathbb{R}^{|S|\times|D|}$ where $|S|$ is the size of the submission archive and $|D|$ is the size of the full collection of terms of all abstracts in the submission archive. For this particular experiment $\mathbf{T}$ is therefore defined as $\mathbf{T} \in \mathbb{R}^{118\times2121}$. The TFIDF term weight vector of each submission can be compared against every other submission's TFIDF term weight vector using the standard cosine similarity function presented in Eq. (5), where $\vec{t}_i$ is the TFIDF term weight vector for submission i. This equation guarantees a symmetrical matrix with a diagonal of 1.0.

$$\mathbf{S}_{\mathrm{t}i,j} = \mathbf{S}_{\mathrm{t}j,i} = \frac{\vec{t}_i \cdot \vec{t}_j}{\|\vec{t}_i\| \cdot \|\vec{t}_j\|} \tag{5}$$

$\mathbf{S_b}$ and $\mathbf{S_t}$ similarities can not be assumed to be parametric[5], therefore the more robust Spearman rank-order correlation, ρ, was adopted to asses correlations. The Spearman ρ correlation between $\mathbf{S}_\mathrm{t}$ and $\mathbf{S}_\mathrm{b}$ was determined to be 0.153. This means that submissions categorized according to a TFIDF analysis of their abstracts and submissions categorized according to the referee bid behavior are in fact positively correlated, though not strongly.

$$\rho = 0.153,\ p < 2.2^{-16}$$

In literature, there exists arguments against judging research similarity according to manuscript term usage (Leydesdorff, 1989, 1997). While TFIDF weighting is an established method for determining document similarity (Baeza-Yates & Ribeiro-Neto, 1999), the construction of $\mathbf{S}_\mathrm{t}$ occurs only at the single term level and therefore, a term may have several conceptual mappings (e.g. *network* in the graph theoretic sense and in the communication infrastructure sense). Thus, $\mathbf{S}_\mathrm{t}$ is inevitably an approximation of the conceptual mapping between the various submissions. However, given that the analyzed abstracts were generated by a small, well-focused community of researchers, such homonyms may not be so prevalent (Leydesdorff, 1989).

5. The bid matrix and referee similarity

This section will overview the experiment as described by Track 2 of Fig. 2. If referees are deemed similar in expertise, as determined by their relative location to one another within the scientific community's co-authorship network, then similar referees should be bidding similarly. To test this hypothesis, two referee similarity matrices are created. The first referee similarity matrix, $\mathbf{R}_\mathrm{b} \in \mathbb{R}^{|R|\times|R|}$, is constructed from the transpose of the modified bid matrix, $\mathbf{B}'^{\mathrm{T}}$. Each referee is compared to each other referee with respect to their bidding behavior. Based on the transpose of the artificial data from Table 2, the same similarity equation used to construct the submission similarity matrix, Eq. (1), can be used to construct the referee similarity matrix presented in Table 12. The next section will present a dendrorgam of $\mathbf{R}_\mathrm{b}$ before discussing the second referee similarity matrix, $\mathbf{R}_\mathrm{g}$.

5.1. Referee similarity and the dendrogram

Given the referee similarity matrix, $\mathbf{R}_\mathrm{b}$, the dendrogram in Fig. 6 can be constructed. Unfortunately, due to privacy issues, the referee names are not provided. What is noticeable from the dendrogram is the collection of

[5] Although $\mathbf{S_b}$ follows a normal distribution ($F = 0.858$), but the similarities of $\mathbf{S_t}$ follow an exponential distribution.

Table 12
Referee similarity determined according to their Hamming distance

Ref	1	2	3	4	5
1	1.0	0.5	0.75	0.0	0.0
2	0.5	1.0	0.2	0.4	0.75
3	0.75	0.2	1.0	0.2	0.0
4	0.0	0.4	0.2	1.0	1.0
5	0.0	0.75	0.0	1.0	1.0

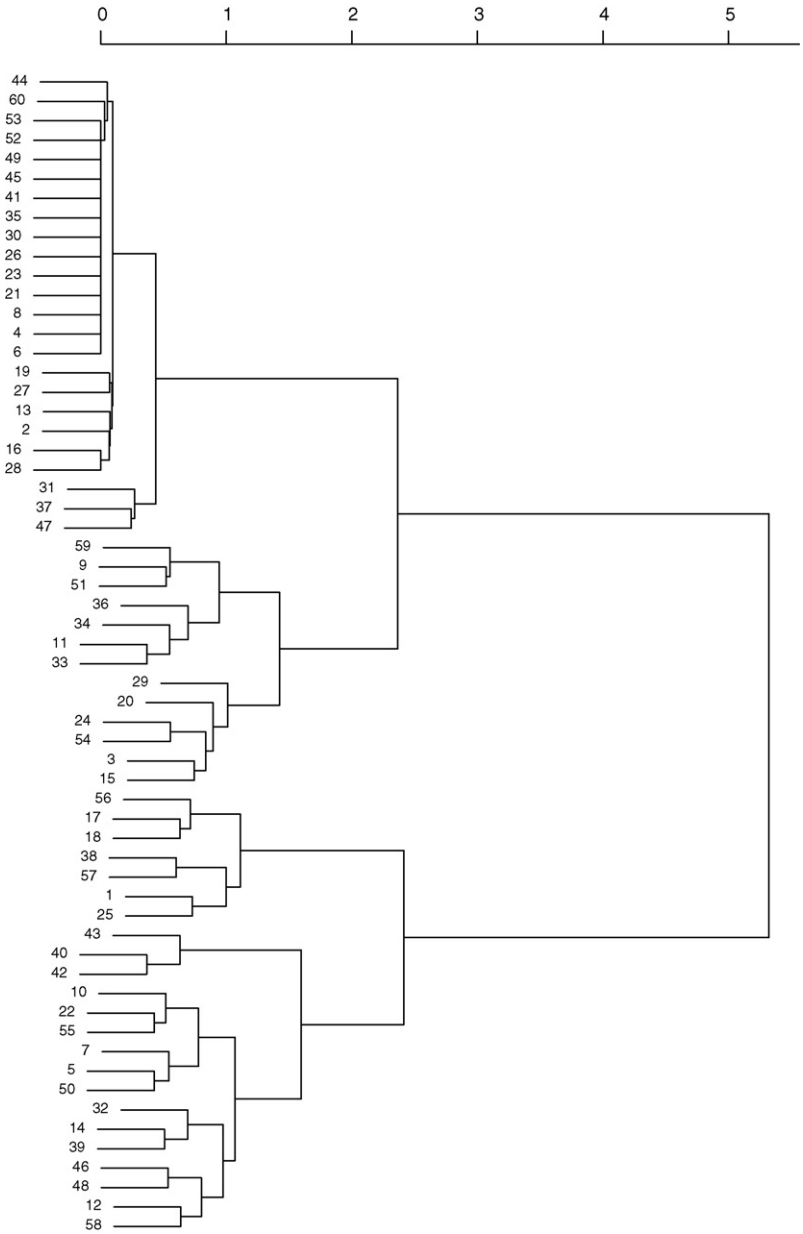


Fig. 6. Referee similarity represented according to a hierarchical cluster.

nearly identical referees on the upper branch. When reviewing the modified bid matrix, $\mathbf{B}'$, it becomes apparent that 19 of the referees stated themselves to be expert in the domain of every submission (excluding their wildcard bids).

5.2. Relative-rank correlation

In order to provide a quantitative evaluation of the similarity of referees with respect to their bidding behavior and their domain of expertise, a relative-rank algorithm within a co-authorship network is computed to determine referee similarity. It has been widely accepted that co-authorship networks represent the relationship of individuals with respect to their domain of expertise (Newman, 2004). The relative-rank algorithm will determine the similarity of each referee with respect to each other referee as defined by their relative location to one another within the greater scientific community's co-authorship network. The similarity of the referees as determined by their relative-rank, $\mathbf{R}_{\mathrm{g}}$, and their similarity as determined by their bid behavior, $\mathbf{R}_{\mathrm{b}}$, can then be correlated. A high correlation means that referees of similar expertise are bidding in a similar manner. A low correlation means that referees of similar expertise are not bidding in a similar manner. The co-authorship network, G, used for this experiment was constructed from the DBLP database as of October 2005. The DBLP co-authorship network has 284,082 nodes (authors) and 2,167,018 edges (co-authorship relationships). This section will first formalize the co-authorship network data structure and relative-rank algorithm before discussing the results.

A co-authorship network is defined by a graph composed of nodes that represent authors and edges that represent a joint publication. Therefore, a co-authorship network is represented by $G = (N, E, W)$, where N is the set of authors in the network, E is the set of edges relating the various authors, and W is the set of weights associated with the strength of tie between any two collaborating authors. Any edge, $e_{i,j}$, connects two authors, n_i and n_j, with a respective weight of $w_{i,j}$. Furthermore, $E \subseteq N \times N$ and $|E| = |W|$. The edge weight between any two authors is determined by Eq. (6), where the summation is over the set of all manuscripts registered with the DBLP, M, expressing a collaboration between authors n_i and n_j, and the function $A(m)$ returns the total number of authors for manuscript m, where $m \in M$ and $w_{i,j} \in \mathbb{R}^+$ (Liu, Bollen, Nelson, & de Sompel, 2005; Newman, 2001).

$$w_{i,j} = w_{j,i} = \sum_{\forall m \in M \text{ authored by } i,j} \frac{1}{A(m) - 1} \tag{6}$$

To provide the reader with an understanding of the relationship between the 2005 JCDL program committee members, a subset of the DBLP co-authorship network which contains the program committee's co-authorship relationships is presented in Fig. 7. Note that this network was not constructed using referee bid data, but from information that is publicly available through the DBLP database. Furthermore, the co-authorship edge weights have been left out to improve readability.

The final analysis to be performed is to rank each of the 60 referees relative to one another so as to construct $\mathbf{R}_{\mathrm{g}} \in \mathbb{R}^{|R| \times |R|}$, Eq. (7). For each referee in the JCDL program committee that provided valid bid data and is located in the DBLP, a similarity value to every other member in the committee was computed using a relative-rank algorithm (sometimes called a 'personalized' rank) (Rodriguez & Bollen, 2005; White & Smyth, 2003)within the DBLP co-authorship network. Since G is a weighted graph, the ranking algorithm actually used in this experiment is the weighted relative-rank implementation described in (Rodriguez & Bollen, 2005).

$$\mathbf{R}_{\mathrm{g}} = \begin{pmatrix} \mathbf{R}_{\mathrm{g}_{R_1,R_1}} & \cdots & \mathbf{R}_{\mathrm{g}_{R_1,R_{|R|}}} \\ \vdots & \ddots & \vdots \\ \mathbf{R}_{\mathrm{g}_{R_{|R|},R_1}} & \cdots & \mathbf{R}_{\mathrm{g}_{R_{|R|},R_{|R|}}} \end{pmatrix} \tag{7}$$

An example of relative-ranking is as follows. Given a network, such as the one displayed in Fig. 7, the relative-rank algorithm would rank *FOX* more strongly to *NELSON* than to *RAY* since there exists a clique relationship between *FOX*, *NELSON*, and their co-authors. This network structure does not exist between *FOX* and *RAY*. Since co-authorship networks relate individuals with respect to similar domains of expertise, the conclusion to be drawn is that the stronger ranking of *FOX* to *NELSON* implies that *FOX* is more related by expertise to *NELSON* than he is to *RAY*. A simpli-

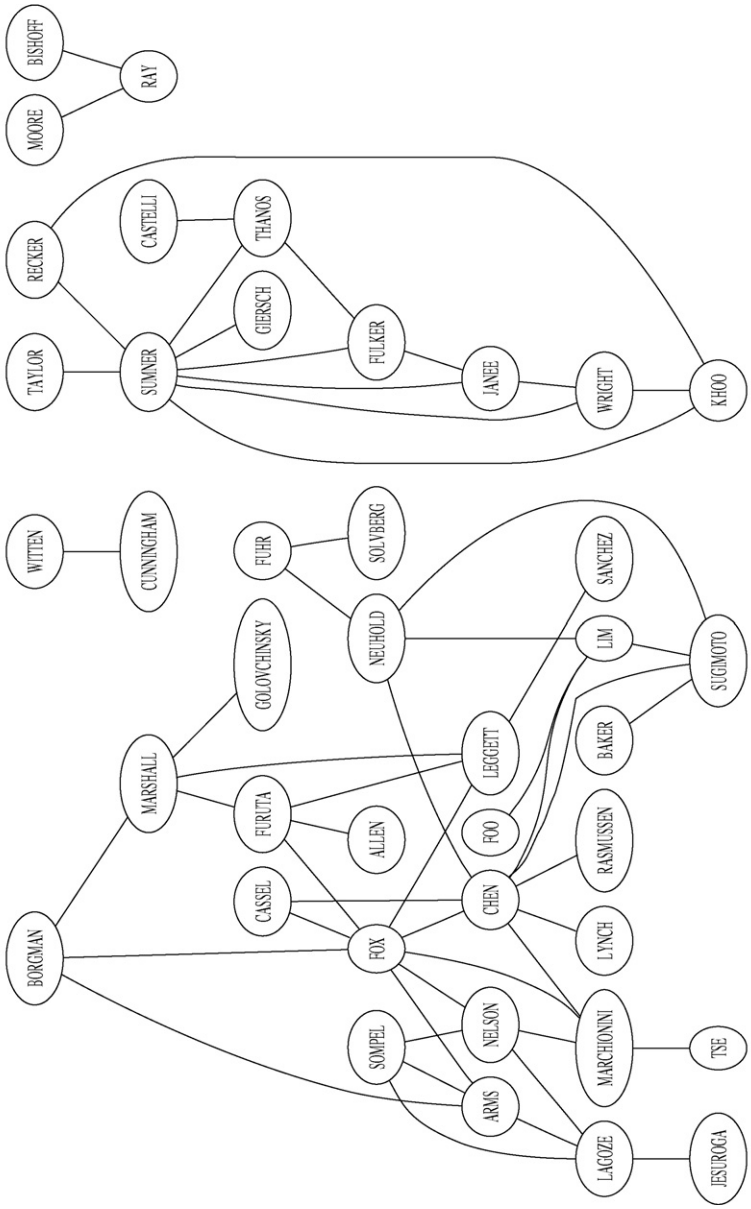


Fig. 7. Subset of the DBLP co-authorship network containing only connected JCDL referees.

fied version of the pseudo-code for constructing $\mathbf{R}_g$, Eq. (7), is presented in Algorithm 1. For a more indepth, and formal, review of relative-rank algorithms for network analysis, refer to (Rodriguez & Bollen, 2005; White & Smyth, 2003).

Algorithm 1. Constructing the referee similarity matrix $\mathbf{R}_g$

```
foreach (n_l ∈ R) do
    foreach (n_j ∈ R) do
        R_g[n_l, n_j] = rank(n_l, n_j);
    end
end
```

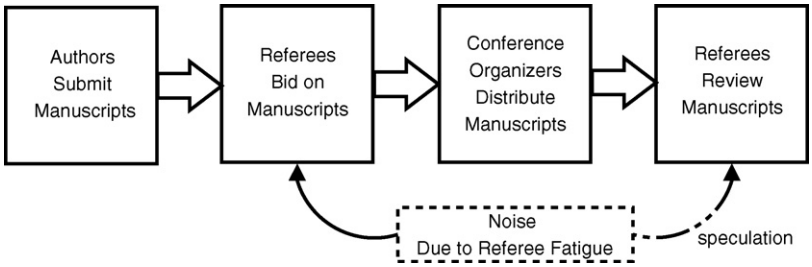


Fig. 8. Human-factor noise in review stages.

Because the similarity values of $\mathbf{R}_g$ are not on an interval scale, the use of a non-parametric correlation test, such as the Spearman rank-order correlation, was warranted. The Spearman rank-order correlation between $\mathbf{R}_b$ and $\mathbf{R}_g$ values was calculated to be 0.101. The positive correlation indicates that referees are bidding with respect to their domain of expertise, but the low correlation may point to other factors.

$$\rho = 0.101, \quad p < 2.2^{-16}$$

Co-authorship networks are not the only data from which author similarity within a domain can be determined. For instance, many authors may publish single authored manuscripts and thus may find themselves not connected within the greater scientific community's co-authorship network. Furthermore, not all collaborations, and therefore similarity of expertise, lead to co-authored publications (Borgman & Furner, 2002) and therefore any assessment of author similarities based on collaboration relationships derived from co-authorship provides only a partial picture of the entire space of existing similarities. Nevertheless, those collaborations that have effectively resulted in co-authorships are positively identified by this method. Using a wider range of bibliographic data may result in an incrementally more complete assessment of author similarity. Unfortunately, such an effort was beyond the scope of this study, but can be included in follow-up studies.

6. Conclusion

This paper provided an exploration of the bidding behavior of the 2005 JCDL program committee. The various analysis techniques used demonstrate that the 2005 JCDL program committee did, in fact, bid for conference submissions with respect to the subject domain of the submission. On the other hand, the strength of this relationship is low and therefore demonstrates that other factors may be involved in referee bidding. One such factor seems to be referee fatigue. With 146 submissions having no bid data and with 19 referees stating themselves to be an expert in the domain of all submissions, human-driven referee bidding in conference settings may not be the most optimal technique for performing conference peer-review. Since bidding is the preliminary component of the manuscript-to-referee matching algorithm, sloppy bidding can have dramatic effects on which referees actually review which submissions, Fig. 8. In general, the stages that follow from the inclusion of noisy data in the peer-review chain can severely effect the quality of the peer-review process. It is speculated that referee fatigue not only influences the bidding and manuscript dissemination stages of the review cycle, but potentially more damaging, fatigued referees could be rejecting acceptable manuscripts or accepting fraudulent or faulty manuscripts in the review stage.

Future work in this area will focus on expanding this study's hypothesis in order to develop a mathematical model of the factors influencing referee bidding. Furthermore, an application of this methodology to bid data from other conferences can help to provide a broader perspective (and confirmation) of the factors influencing submission bidding in the peer-review process.

Acknowledgments

This research could only have been conducted with the help of the 2005 JCDL program chairs (Mary Marlino, Tamara Sumner, Frank Shipman) and steering committee (Erich Neuhold). Furthermore, the DBLP has once again provided the authors of this paper a thorough dataset for exploring the structure of science. The authors would like to thank the producers of R Statistics, GraphViz, Dia, and LaTeXfor their freely available software. Finally, Xiaoming Liu

and Karin Verspoor contributed by reviewing drafts of this manuscript. This research was financially supported by the Los Alamos National Laboratory.

References

Baeza-Yates, R. A., & Ribeiro-Neto, B. A. (1999). *Modern Information Retrieval*. ACM Press/Addison-Wesley.

Bence, V., & Oppenheim, C. (2004). The influence of peer review on the research assessment exercise. *Journal of Information Science*, *30*(4), 347–368.

Borgman, C., & Furner, J. (2002). Scholarly communication and bibliometrics. *Annual Review of Information Science and Technology*, *36*, 3–72.

Eisenhart, M. (2002). The paradox of peer review: Admitting too much or allowing too little? *Research in Science Education*, *32*(2), 241–255.

El-Munchid, H. (2001). Evaluation of peer review in biomedical publications. *Annals of Saudi Medicine*, *21*, 5–6.

Evans, P. (1995). The peer-review process. *Literati Newsline: Special Issue for Authors and Editors*, 2.

Hamming, R. W. (1950). Error-detecting and error-correcting codes. *Bell System Technical Journal*, *29*(2), 147–160.

Kassirer, J. P., & Campion, E. W. (1994). Peer review: Crude and understudied, but indispensable. *The Journal of the American Medical Association*, *272*, 96–97.

Lance, G., & Williams, W. (1967). A general theory of classificatory sorting strategies i. clustering systems. *The Computer Journal*, *10*(3), 271–277.

Leydesdorff, L. (1989). Words and co-words as indiactors of intellectual organization. *Research Policy*, *18*, 209–223.

Leydesdorff, L. (1997). Why words and co-words cannot map the development of science. *Journal of the American Society for Information Science*, *45*(5), 418–427.

Liu, X., Bollen, J., Nelson, M. L., & de Sompel, H. V. (2005). Co-authorship networks in the digital library research community. *Information Processing and Management*, *41*, 1462–1480.

Newman, M. E. J. (2001). Scientific collaboration networks. I. Network construction and fundamental results. *Physical Review*, 64.

Newman, M. E. J. (2004). Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Science*, 5200–5205.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program: Automated Library and Information Systems*, *14*(3), 130–137.

Rodriguez, M. A., & Bollen, J. (2005). *Simulating network influence algorithms using particle-swarms: Pagerank and pagerank-priors*. Technical report. Los Alamos National Laboratory [LAUR-05–6469].

Salton, G. (1998). Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, *24*(5), 513–523.

Shannon, C. E. (1948). A matematical theory of communication. *Bell System Technical Journal*, *27*, 379–423.

Sumner, T. (2005). Report on the fifth ACM/IEEE joint conference on digital libraries—cyberinfrastructure for research and education. *D-Lib Magazine*, 11(7/8).

Wei, J. C., Hartvigsen, D., & Czuchlewski, R. (1999). The conference paper-reviewer assignment problem. *Decision Science*, 30(3).

White, S., & Smyth, P. (2003). Algorithms for estimating relative importance in networks. In *KDD '03: Proceedings of the ninth ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 266–275). New York, NY, USA: ACM Press.

Available online at www.sciencedirect.com
ScienceDirect
Journal of Informetrics 1 (2007) 35–46
Journal of INFORMETRICS
An International Journal
www.elsevier.com/locate/joi

Measuring quality of similarity functions in approximate data matching

Roberto da Silva*, Raquel Stasiu[1], Viviane Moreira Orengo, Carlos A. Heuser

UFRGS, Instituto de Informática, Porto Alegre, Brazil

Received 3 July 2006; received in revised form 23 August 2006; accepted 5 September 2006

Abstract

This paper presents a method for assessing the quality of similarity functions. The scenario taken into account is that of approximate data matching, in which it is necessary to determine whether two data instances represent the same real world object. Our method is based on the semi-automatic estimation of optimal threshold values. We propose two methods for performing such estimation. The first method is an algorithm based on a reward function, and the second is a statistical method. Experiments were carried out to validate the techniques proposed. The results show that both methods for threshold estimation produce similar results. The output of such methods was used to design a grading function for similarity functions. This grading function, called *discernability*, was used to compare a number of similarity functions applied to an experimental data set.

Keywords: Approximate data matching; Similarity functions; Retrieval evaluation

1. Introduction

The process of approximate data matching aims at defining whether two data instances (strings, tuples, trees, ...) represent the same real world object. This process appears in several data management applications such as *approximate querying* and *data integration*. In approximate querying, the problem is to find database instances that represent the same data instance given as a query. In data integration, the aim is to assess whether two data instances originating from different sources represent the same real world object. Approximate data matching usually relies on the use of a *similarity function*. A similarity function $f(v_1, v_2) \mapsto s$ assigns a score s to a pair of data values v_1 and v_2. These values are considered to be representing the same real world object if s is greater then a given *threshold* t. There is a wide range of similarity functions, from very simple string matching functions, like Levenshtein's edit distance (Hall & Dowling, 1980; Levenshtein, 1966; Navarro, Baeza-Yates, Sutinen, & Tarhio, 2001), to functions specific to XML trees (Dorneles, Heuser, Lima, da Silva, & de Moura, 2004). Generally speaking, similarity functions are imperfect and the quality of their results will depend on the specific data set being matched.

The use of similarity functions in approximate data matching poses two problems. The first is to determine the threshold value that should be used. The difficulty in this case arises from the fact that the distribution of score values

* Corresponding author. Tel.: +55 51 33167772; fax: +55 51 3316 7308.
E-mail addresses: rdasilva@inf.ufrgs.br (R. da Silva), rkstasiu@inf.ufrgs.br (R. Stasiu), vmorengo@inf.ufrgs.br (V.M. Orengo), heuser@inf.ufrgs.br (C.A. Heuser).
[1] On leave from PUC-PR and UTFPR.

1751-1577/$ – see front matter
doi:10.1016/j.joi.2006.09.001

obtained by one similarity function may be completely different from the distribution obtained by another. It may even vary when the same similarity function is applied to different data sets.

The second problem is how to measure if a similarity function is more adequate for a specific data set then another. Existing approaches for the evaluation of similarity functions (Bilenko, Mooney, Cohen, Ravikumar, & Fienberg, 2003; Cohen, 2003) are based on the recall/precision curve, a classical Information Retrieval (IR) quality measure (Salton, 1989). Recall/precision curves are useful to express the ability of a similarity function in ranking the results of matches. However, they are not suitable for expressing how efficient similarity functions are in telling apart relevant from irrelevant matches.

In this paper, we propose a quality measure specifically designed for similarity functions in the context of data matching. As a byproduct, our approach also produces a threshold value that may be interpreted as the "best" one for a given similarity function, when considering a specific data set. Here, "best" means a threshold value that minimizes false positives and false negatives with respect to an answer set.

The remainder of this paper is organized as follows: Section 2 proposes two methods for threshold definition; Section 3 presents experiments that evaluate the proposed methods; Section 4 proposes a function called *discernability* to assess the quality of similarity functions and applies it to compare several similarity functions; Section 5 presents a summary and the conclusions.

2. Process of threshold definition

For most applications, the process of threshold definition is left to the user who must choose an arbitrary value to be applied to one or more queries. If the threshold chosen is too high, there is a risk of not retrieving any results. On the other hand, if the chosen threshold is too low, many irrelevant items will be retrieved. This problem is aggravated by the fact, mentioned in the introduction, that the distribution of score values may vary significantly from one similarity function to another. As a result, the definition of a threshold is generally a trial and error process, in which the user has to test a number of different values until the result is satisfactory.

In this section, we propose two semi-automatic methods for the calculation of threshold values for a given similarity function. The output of these methods is an interval of threshold values $[t_{\text{best}}^{\min}, t_{\text{best}}^{\max}]$ that provides optimum results. By optimum we mean a threshold value that maximizes the number of cases in which $s_{\text{irrel}} \leq t_{\text{best}} \leq s_{\text{rel}}$, where s_{rel} is the lowest score for a relevant item, s_{irrel} is the highest score for an irrelevant item. In what follows, we explain how to calculate these scores. The process of threshold definition should be guided by two premises: (i) minimize false positives and (ii) minimize false negatives.

Both methods rely on a sampling process that takes values from a pre-existing *collection* V of data values (v) to be compared by a similarity function. These sample values are then used as queries against V in order to collect knowledge about how the score values are distributed.

More specifically, the sampling process is as follows: A *sample* $Q \subseteq V$ is taken from the collection. Each element $q \in Q$ is used as a query object against the collection V. The similarity between the query and each element of the database is calculated using a given similarity function:

$$L : (Q \subseteq V) \times V \to \mathbb{R}^+.$$

So, for each $q \in Q$ we define the set:

$$R_q = \{s \in \mathbb{R}^+ / s = L(q, v),\ v \in V\}.$$

Here, R_q induces an order $\prec$ on V, defined by relation:

$$v, w \in V, \quad v \prec w \Leftrightarrow L(q, v) < L(q, w),$$

then the values in V are ranked in decreasing order of similarity to the query value q.

Next, a human expert labels each element of the ranking as `relevant (rel)`, if the data value is considered to represent the same real world object as the query q or as `irrelevant (irrel)` otherwise. Defining

$$v_q(\text{rel}) = \min\{v / v \text{ is relevant}\}, \qquad v_q(\text{irrel}) = \max\{v / v \text{ is irrelevant}\},$$

if $n = |Q|$, $q \in Q$ we note k the index of q such that $k \in [1, n]$.

Table 1
Example of similarity ranking

Score	Data item	Relevance
1.0000	Journal of Informetrics	Relevant
0.8636	Jrnl of Infometrics	Relevant
0.7391	J. of Informetrics	Relevant
0.1304	Informetrics Journal	Relevant
0.1304	JOI	Relevant
0.1250	Decision Support Systems	Irrelevant
0.0869	TODS	Irrelevant
0.0869	SIGMOD	Irrelevant
0.0434	TKDE	Irrelevant

This labelling enables us to identify two important points in the ranking: $s^L_{\text{rel}}(k) = L(q, v_q(\text{rel}))$, which is the lowest score corresponding to a relevant item and $s^L_{\text{irrel}}(k) = L(q, v_q(\text{irrel}))$, which is the highest score attained by an irrelevant item. Those values are used by both methods for threshold definition proposed in this paper. Notice that for some queries $s^L_{\text{irrel}}(k)$ could be greater than $s^L_{\text{rel}}(k)$. Such a situation indicates that the similarity function has failed to separate relevant from irrelevant items.

Example. Consider a database containing titles of computing science journals. The object "Journal of Informetrics" is represented in five different forms, namely: "Journal of Informetrics", "J. of Informetrics", "JOI", "Informetrics Journal", "Jrnl of Infometrics". Supposing that the database contains nine data items, the ranking generated by the edit distance function is shown in Table 1. According to this ranking, the lowest score of a relevant item is $s_{\text{rel}} = 0.1304$ and the highest score of an irrelevant item is $s_{\text{irrel}} = 0.1250$.

2.1. Reward function algorithm

A function to measure how good a threshold value is in separating relevant from irrelevant items can be defined by the simple formula below:

$$f^L(n, t) = \sum_{k=1}^{n} d(s^L_{\text{rel}}(k), s^L_{\text{irrel}}(k)) \tag{1}$$

where L is the similarity function used; n the number of queries (sample size); t is the threshold being analyzed; $d(\cdot, \cdot)$ measures how adequate $s^L_{\text{rel}}(k)$ and $s^L_{\text{irrel}}(k)$ are with the threshold t, such that:

$$d(s^L_{\text{rel}}(k), s^L_{\text{irrel}}(k)) = R^t_{\text{rel}}(k) + R^t_{\text{irrel}}(k) \tag{2}$$

with

$$R^t_{\text{rel}}(k) = \begin{cases} 1 & \text{if } s^L_{\text{rel}}(k) > t \\ -1 & \text{else } s^L_{\text{rel}}(k) \le t \end{cases} \quad \text{and} \quad R^t_{\text{irrel}}(k) = \begin{cases} -1 & \text{if } s^L_{\text{irrel}}(k) \ge t \\ 1 & \text{else } s^L_{\text{irrel}}(k) < t \end{cases} \tag{3}$$

According to these equations, the optimal threshold t_{best} (or more precisely the interval for the optimal threshold), which reaches the maximum value on the function $f^L(n, t)$, can be defined as

$$f^L_{\max} = \max_{t \in [t_{\min}, t_{\max}]} \{f^L(n, t)\}. \tag{4}$$

where $t_{\min}$ and $t_{\max}$ represent the limits of the threshold interval to be tested.

Algorithm 1 shows a description of BestThresh, which determines t_{best}. The inputs for this algorithm are: (i) the number of queries (n); (ii) the limits of the threshold interval to be tested ($t_{\min}$ and $t_{\max}$); (iii) the lowest similarity score achieved by a relevant item for the query k, denoted by $s^L_{\text{rel}}(k)$; (iv) the highest score achieved by an irrelevant item for the same query k, denoted by $s^L_{\text{irrel}}(k)$; (v) the numerical precision (h) on which the algorithm should operate. The algorithm produces two outputs: the interval $[t^{\min}_{\text{best}}, t^{\max}_{\text{best}}]$ in which the optimal threshold

(t_{best}) lies; and its associated f_{max}, which is the number of points achieved by that threshold interval. The reason for the output of the algorithm being an interval and not a single value is that a number of threshold values, in sequential order, can achieve f_{max}. The lowest and the highest values are then used as limits of the interval. Ways in which f_{max} could be used in the evaluation of the quality of the similarity function will be discussed in Section 4.

The limits t_{min} and t_{max} are, respectively, the smallest and the largest similarity scores from the ranking generated by the similarity function. The numerical precision, denoted by h, is calculated by the formula $h = (t_{max} - t_{min})/n_{div}$, where n_{div} is the number of divisions we want to make on the interval $[t_{min}, t_{max}]$. This way, each threshold t to be tested by the algorithm is obtained by $t_i = t_{min} + ih$, where $i = 0, \ldots, n_{div}$.

The algorithm works as follows: each threshold t between t_{min} and t_{max} is tested for each query. The test consists in comparing t with s_{rel} and s_{irrel}. The number of points achieved by each threshold t is computed according to Eqs. (3) and (4). The highest number of points achieved by a threshold (f_{max}), which is initialized at the beginning of the algorithm with the smallest value possible, is then found. Once f_{max} is established, the algorithm finds the interval in which all threshold values achieve f_{max}.

Algorithm 1. BestThresh

Input: $n, t_{min}, t_{max}, s^L_{rel}(k), s^L_{irrel}(k), k = 1 \ldots n, h$
Output: $t^{min}_{best}, t^{max}_{best}, f_{max}$
$f_{max} = -2n$;
$n_{div} = (t_{max} - t_{min})/h$
for (a) $i = 0, \ldots, n_{div}$ **do**
  $t = t_{min} + ih$;
  $f(t) = 0$;
  for (b) $k = 1, \ldots, n$ **do**
    $d = 0$;
    if ($s^L_{rel}(k) > t$) **then**
      $d = d + 1$;
    else
      $d = d - 1$;
    end if
    if ($s^L_{irrel}(k) < t$) **then**
      $d = d + 1$;
    else
      $d = d - 1$;
    end if
    $f(t) = f(t) + d$;
  end for (b)
  if ($f(t) \geq f_{max}$) **then**
    $f_{max} = f(t)$;
  end if
end for (a)
$t = t_{min}$
while ($f(t) \neq f_{max}$) **do**
  $t = t + h$
end while
$t^{min}_{best} = t$
$t = t_{max}$
while ($f(t) \neq f_{max}$) **do**
  $t = t - h$
end while
$t^{max}_{best} = t$
if $f_{max} < 0$ **then**
  aux $= t^{max}_{best}$
  $t^{max}_{best} = t^{min}_{best}$
  $t^{min}_{best} =$ aux
end if
Write "the best threshold is in the interval" $[t^{min}_{best}, t^{max}_{best}]$

2.2. Bivariate normal distribution

In this section, we explore the approach of a bivariate normal distribution (Spiegel, 1992; Weisstein, 2004) to find t_{best}. This method is based on statistics and tries to maximize the probability of finding a threshold that minimizes false positives and false negatives.

Let us consider the probability density function (PDF) for s^L_{irrel} and s^L_{rel}, denoted by $P(s^L_{\text{irrel}})$ and $P(s^L_{\text{rel}})$, respectively. Considering a sample of size n, the experimental mean values are computed as $\langle s^L_{\text{irrel}}\rangle = (1/n)\sum_{k=1}^{n} s^L_{\text{irrel}}(k)$ and $\langle s^L_{\text{rel}}\rangle = (1/n)\sum_{k=1}^{n} s^L_{\text{rel}}(k)$, and the respective standard deviation $\sigma(s^L_{\text{irrel}}) = \sqrt{[1/(n-1)]\sum_{k=1}^{n}[s^L_{\text{irrel}}(k) - \langle s^L_{\text{irrel}}\rangle]^2}$, $\sigma(s^L_{\text{rel}}) = \sqrt{[1/(n-1)]\sum_{k=1}^{n}[s^L_{\text{rel}}(k) - \langle s^L_{\text{rel}}\rangle]^2}$.

Given the means and the standard deviations, we can calculate the distributions for $P(s_{\text{rel}})$ and $P(s_{\text{irrel}})$, which would approximately be:

$$P(s^L_{\text{irrel}}) = \frac{1}{\sqrt{2\pi\sigma^2(s^L_{\text{irrel}})}} \exp\left[-\frac{1}{2}\left(\frac{s^L_{\text{irrel}} - \langle s^L_{\text{irrel}}\rangle}{\sigma(s^L_{\text{irrel}})}\right)^2\right],$$

$$P(s^L_{\text{rel}}) = \frac{1}{\sqrt{2\pi\sigma^2(s^L_{\text{rel}})}} \exp\left[-\frac{1}{2}\left(\frac{s^L_{\text{rel}} - \langle s^L_{\text{rel}}\rangle}{\sigma(s^L_{\text{rel}})}\right)^2\right] \tag{5}$$

Since there is a correlation between $s^L_{\text{irrel}}(k)$ and $s^L_{\text{rel}}(k)$, the joint distribution $P(s^L_{\text{irrel}}, s^L_{\text{rel}})$ is not necessarily the product $P(s^L_{\text{irrel}}) \cdot P(s^L_{\text{rel}})$. Therefore, in order to calculate the joint PDF, we need to take the correlation coefficient ρ into consideration. The formula for the joint PDF is given below:

$$P(s^L_{\text{irrel}}, s^L_{\text{rel}}) = \frac{1}{2\pi\sigma(s^L_{\text{rel}})\sigma(s^L_{\text{irrel}})\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{s^L_{\text{irrel}} - \langle s^L_{\text{irrel}}\rangle}{\sigma(s^L_{\text{irrel}})}\right)^2 + \left(\frac{s^L_{\text{rel}} - \langle s^L_{\text{rel}}\rangle}{\sigma(s^L_{\text{rel}})}\right)^2 - 2\rho\left(\frac{s^L_{\text{irrel}} - \langle s^L_{\text{irrel}}\rangle}{\sigma(s^L_{\text{irrel}})}\right)\left(\frac{s^L_{\text{rel}} - \langle s^L_{\text{rel}}\rangle}{\sigma(s^L_{\text{rel}})}\right)\right]\right\}, \tag{6}$$

where ρ is the correlation coefficient defined by the formula:

$$\rho = \frac{\langle s^L_{\text{rel}} s^L_{\text{irrel}}\rangle - \langle s^L_{\text{rel}}\rangle\langle s^L_{\text{irrel}}\rangle}{\sigma(s^L_{\text{rel}})\sigma(s^L_{\text{irrel}})}, \tag{7}$$

that assumes values in the interval $[-1, 1]$, where $|\rho| \sim 1$ denotes correlated data and $|\rho| \sim 0$ denotes that in our data, s^L_{rel} and s^L_{irrel} are independent random variables.

For determining t_{best}, it is sufficient to find the value of t that yields the maximum:

$$F(t) = P(s^L_{\text{irrel}}\langle t, s^L_{\text{rel}}\rangle t) = \frac{1}{2\pi\sigma(s^L_{\text{rel}})\sigma(s^L_{\text{irrel}})\sqrt{1-\rho^2}} \int_{-\infty}^{t}\int_{t}^{\infty} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{s^L_{\text{irrel}} - \langle s^L_{\text{irrel}}\rangle}{\sigma(s^L_{\text{irrel}})}\right)^2 + \left(\frac{s^L_{\text{rel}} - \langle s^L_{\text{rel}}\rangle}{\sigma(s^L_{\text{rel}})}\right)^2 - 2\rho\left(\frac{s^L_{\text{irrel}} - \langle s^L_{\text{irrel}}\rangle}{\sigma(s^L_{\text{irrel}})}\right)\left(\frac{s^L_{\text{rel}} - \langle s^L_{\text{rel}}\rangle}{\sigma(s^L_{\text{rel}})}\right)\right]\right\} \mathrm{d}s^L_{\text{irrel}}\,\mathrm{d}s^L_{\text{rel}} \tag{8}$$

That is, we are trying to find the value of t that maximizes the probability of t being simultaneously greater than s^L_{irrel} and less than s^L_{rel}. A simple algorithm was implemented to compute $F(t)$, in order to discover the value of t_{best} for each similarity function used in the experiments.

3. Experiments

In this section, we describe the experiments we carried out in order to evaluate the two threshold definition methods proposed in Section 2. We collected titles for 18 scientific papers and manually edited them (i.e. adding, replacing, removing and/or swapping characters or words) to simulate possible typing errors. In total 150 paper titles were

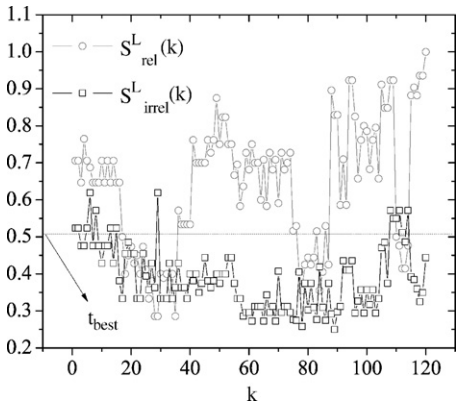


Fig. 1. Lowest relevant score and highest irrelevant score as function of the kth query for function L (edit distance).

generated. A sample of 120 items was picked and used as queries against the database. The similarity between the each query and the documents was calculated using the Edit distance function (Hall & Dowling, 1980; Levenshtein, 1966; Navarro et al., 2001). This function calculates the minimum number of character insertions, deletions and replacements necessary to make two strings equal.

As mentioned in Section 2, a human expert labelled all returned items in the ranked list as relevant or irrelevant. Based on this labelling, the values for $s^L_{\text{rel}}(k)$ and $s^L_{\text{irrel}}(k)$ were obtained. Fig. 1 shows a plot of $s^L_{\text{rel}}(k)$ and $s^L_{\text{irrel}}(k)$ as a function of the kth query, for a sample of 120 queries (n).

3.1. Experiments using the BestThresh algorithm

Considering a number from $k = 1$ to n queries and a precision of $h = (t_{\max} - t_{\min})/n_{\text{div}} = 0.001$, where $t_{\max} = 1$, $t_{\min} = 0$ and $n_{\text{div}} = 1000$, we run the BestThresh algorithm evaluating each threshold calculated by the formula $t_i = t_{\min} + ih$, where $i = 1, \ldots, n_{\text{div}}$. The results produced by the algorithm indicate that t_{best} lies in the interval $I = [0.524, 0.529]$. All threshold values within this interval have achieved $f_{\max} = f^{\text{edit}}(n, t) = 154$. Thus, whichever value belonging to I would be a suitable threshold for performing a search by chance using this particular similarity function. Fig. 2 shows a plot of $f^{\text{edit}}(n, t)$ as a function t.

Notice that the values of $f^{\text{edit}}(n, t)$ are distributed symmetrically as function of t. The continuous curve in Fig. 2 is a normal fit for our data, which gives us an exact notion of how our values are distributed.

A "robustness" test can be performed to assess how t_{best} behaves as the sample size (n) grows. Intuitively, t_{best} should converge to a constant value t^{∞}_{best} when extrapolating $k \to n$. Fig. 3 shows that t_{best} stabilizes as the number of queries approaches 120.

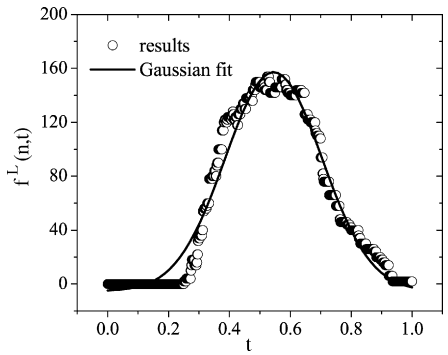


Fig. 2. Plot of $f^{\text{edit}}(n, t)$ as function of t for the edit distance function.

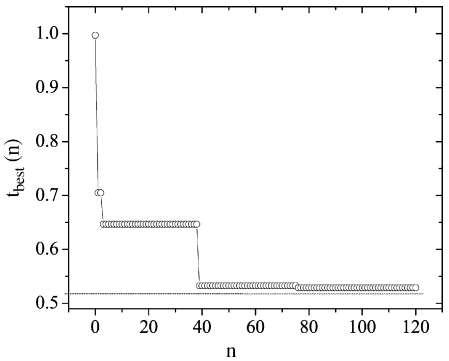


Fig. 3. Evolution of t_{best} as function of sample size. The plot clearly shows that t_{best} converges to the values in the interval $I = [0.524, 0.529]$ as $n \to \infty$.

3.2. Results using bivariate normal distribution

We calculated the parameters $(\langle s^L_{irrel}\rangle, \sigma^2(s^L_{irrel}))$ and $(\langle s^L_{rel}\rangle, \sigma^2(s^L_{rel}))$ for the sample queries using the same similarity function applied in the previous subsection (edit distance). Histograms for the values for s^L_{irrel} and s^L_{rel} were also computed. Fig. 4 shows how s^L_{irrel} and s^L_{rel} are distributed around the mean values $\langle s^L_{irrel}\rangle$ and $\langle s^L_{rel}\rangle$. The continuous curves in these plots denote the normal fits. Calculating the correlation $\rho = 0.022$, we obtained the bivariate normal, shown in Fig. 5.

The numerical software Maple was used to calculate $F(t)$ specified by Eq. (8), spanning t in the interval $[0, 1]$ to find the value of t that yields the maximum $F(t)$. We present our results on Fig. 5 as a plot of $F(t)$ as a function of t. Notice that the probability $F(t)$ is approximately a normal PDF once that the continuous curve is a normal PDF fit. The figure shows that the most likely value for t_{best} is 0.515. This value was obtained by our statistical method

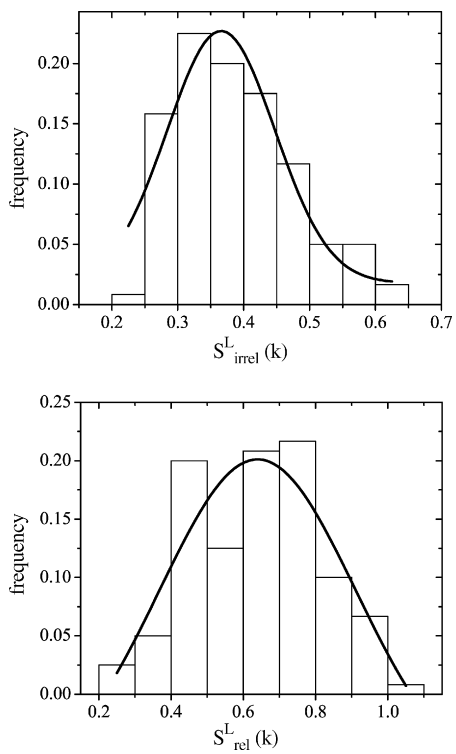


Fig. 4. Histograms for the values of s^L_{irrel} and s^L_{rel}. The continuous curves are the gaussian fits for these histograms. A correlation coefficient between the two distributions can be determined.

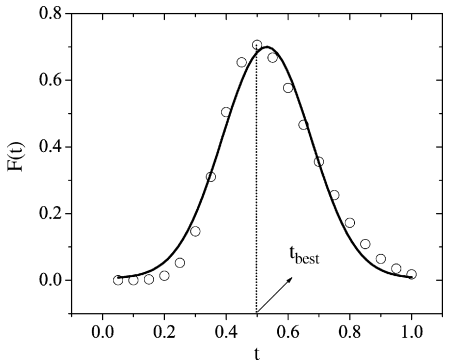


Fig. 5. Plot $F(t) \times t$. The most likely value for t_{best} is the value of t corresponding to the highest value of $F(t)$.

considering a precision of $h = 0.001$. Recall that the value for t_{best} calculated by the BestThresh algorithm was in the interval $I = [0.524, 0.529]$. This shows that both methods are in agreement.

4. Evaluating similarity functions

The aim of this section is to use our two methods for threshold definition to evaluate the quality of different similarity functions. One similarity function can be considered better than another if it provides better separation of relevant and irrelevant data items returned in response to a query. According to our approach, a similarity function that has a higher $f_{\max}$ is considered better than another function that has a smaller $f_{\max}$. Also, the size of the range of the interval for t_{best} is another indicator of the quality of the function. Given that a good similarity function should place relevant and irrelevant items far apart in the ranking, the larger the interval, the better. Section 4.1 proposes a function that evaluates the quality of a similarity function for a specific data set. Section 4.2 applies the proposed discernability to assess a number of similarity functions.

4.1. The discernability function

Below we define a method for assessing the quality of a similarity function. We named the proposed function *discernability* as it refers to the ability of the similarity function in discerning relevant from irrelevant items. Discernability takes two aspects into consideration: (i) how well the similarity function separates relevant from irrelevant items; (ii) how far apart in the ranking the similarity function places relevant and irrelevant items. The first aspect is given by the maximum number of points ($f_{\max}$) calculated by the BestThresh algorithm. The second aspect can be calculated by taking the difference between $t_{\text{best}}^{\max}$ and $t_{\text{best}}^{\min}$. Discernability also defines two coefficients c_1 and c_2 which allow the user to express the importance given to each of the two aspects considered. For the experiments described in this paper we gave the same importance to c_1 and c_2 using $c_1 = c_2 = 1$. The values produced by the discernability will be in the interval $[-1, 1]$.

$$discernability^{L}(t_{\text{best}}^{\min}, t_{\text{best}}^{\max}, f_{\max}) = \frac{c_1}{c_1 + c_2}(t_{\text{best}}^{\max} - t_{\text{best}}^{\min}) + \frac{c_2}{c_1 + c_2} \cdot \frac{f_{\max}}{2n} \tag{9}$$

In order to assess whether the threshold values calculated by our two methods are plausible, we define two measures for computing the theoretical confidence interval, considering the distribution of threshold values. In this case the average value of t is given by

$$\langle t \rangle = \frac{\sum_{i=1}^{n} t_i F(t_i)}{\sum_{i=1}^{n} F(t_i)} \tag{10}$$

and the respective uncertainty associated

$$\sigma_t = \sqrt{\frac{\sum_{i=1}^{n} t_i^2 F(t_i)}{\sum_{i=1}^{n} F(t_i)} - \langle t \rangle^2} \tag{11}$$

4.2. Experiments

The same data used in Section 3 was used to compare the performance of eight similarity functions. Below we list each function together with a brief description.

- **Edit distance** (Hall & Dowling, 1980; Levenshtein, 1966; Navarro et al., 2001). As mentioned in the previous section, this function computes the minimum number of changes (insertions,deletions and replacements) that are necessary to make two strings equal.
- **Acronyms** (Dorneles et al., 2004). This function is useful for matching acronyms to their unabbreviated form, e.g. matching "JOI" to "Journal of Informetrics".
- **Guth** (Guth, 1976). This function is designed for matching proper nouns.
- **Jaccard** (Jaccard, 1912). This simple function states that the similarity between s_1 and s_2 is given by $(s_1 \cap s_2) \div (s_1 \cup s_2)$.
- **Jaro** (Jaro, 1989). This is a function based on the number and order of common characters between two strings.
- **JaroWinkle**r (Winkler, 1999). This is a variant of the Jaro function that emphasizes matches in the first few characters.
- **N-gram** (Navarro et al., 2001). The similarity score is calculated based on the number of characters that are in the same position in each gram. For the experiments described in this section, we used $n = 3$.
- **TF-IDF** (Salton & McGill, 1983). The acronym stands for term-frequency inverse document frequency. This is widely used in IR as a weighting scheme in order to give more importance to less frequent words. For string matching, TF is the frequency of the term in the string and IDF can be computed using the entire collection of strings to be matched.

In addition to the real similarity functions above, we tested three artificial ones:

- **Optimal**. The perfect similarity function should correctly separate relevant and irrelevant items and place them as far as possible in the ranking, i.e. for all queries $s_{\text{rel}} = 1$ and $s_{\text{irrel}} = 0$.
- **NoneRetrieved**. Function that calculates a similarity score of zero between the query and all data items, no matter whether they are relevant or not. In this case for all queries $s_{\text{rel}} = s_{\text{irrel}} = 0$.
- **WorstPossible**. The worst function places all non-relevant items higher than the relevant ones in the ranking, i.e. for all queries $s_{\text{rel}} = 0$ and $s_{\text{irrel}} = 1$.

A precision of $h = 0.001$ was used for the computation of the interval $[t_{\text{best}}^{\min}, t_{\text{best}}^{\max}]$ by the BestTresh algorithm. The limits of this interval were used to calculate the *discernability* for the similarity function. t_{best} was calculated by the bivariate normal distribution.

The results are shown in Table 2. The second column of the table presents the values for $f_{\max}$, which represent the number of points achieved by t_{best} for a given function. The third column displays the results for *discernability*. The fourth column shows the interval for t_{best} calculated by the BestThresh algorithm. The fifth column contains the most

Table 2
Comparison among different similarity functions

Function	$f_{\max}$	Discernability	$[t_{\text{best}}^{\min}, t_{\text{best}}^{\max}]$	t_{best}	Confidence interval
Jaro–Winkler	184	0.4048	[0.768, 0.811]	0.791	[0.716, 0.887]
Jaro	178	0.3713	[0.755, 0.756]	0.753	[0.703, 0.888]
Acronyms	158	0.3616	[0.601, 0.666]	0.592	[0.455, 0.781]
Edit distance	154	0.3233	[0.524, 0.529]	0.515	[0.404, 0.697]
N-gram	134	0.2851	[0.576, 0.588]	0.553	[0.440, 0.723]
Guth	38	0.0821	[0.905, 0.911]	0.801	[0.686, 0.896]
Jaccard	16	0.0468	[0.401, 0.428]	0.301	[0.169, 0.428]
TFIDF	16	0.0438	[0.578, 0.599]	0.442	[0.273, 0.614]
Optimal*	240	0.9999	[0.001, 0.999]	–	–
NoneRetrieved*	0	0.0000	[0.000, 0.000]	–	–
WorstPossible*	−240	−0.9999	[0.999, 0.001]	–	–

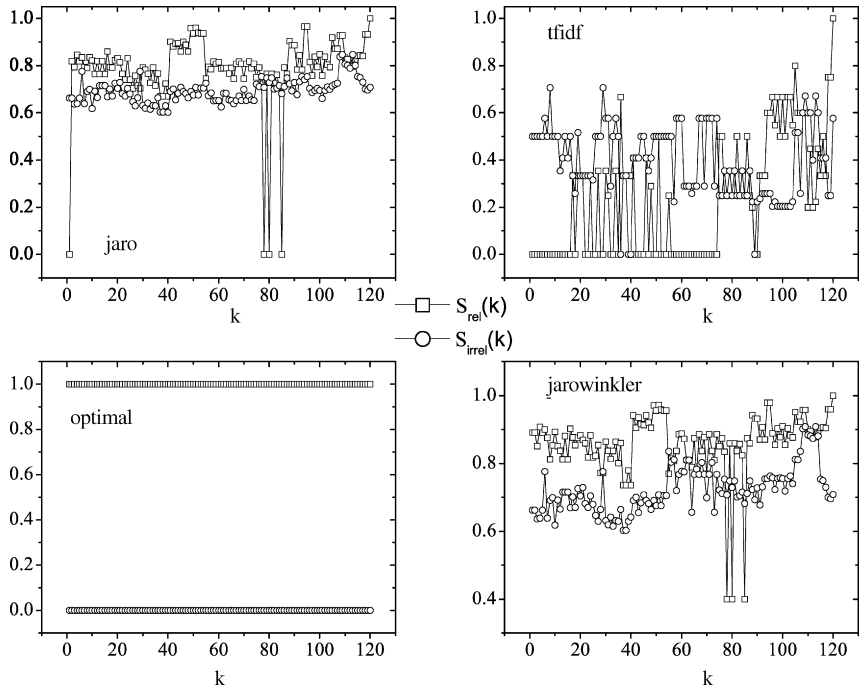


Fig. 6. Distribution of S_{rel} and S_{irrel} for different similarity functions.

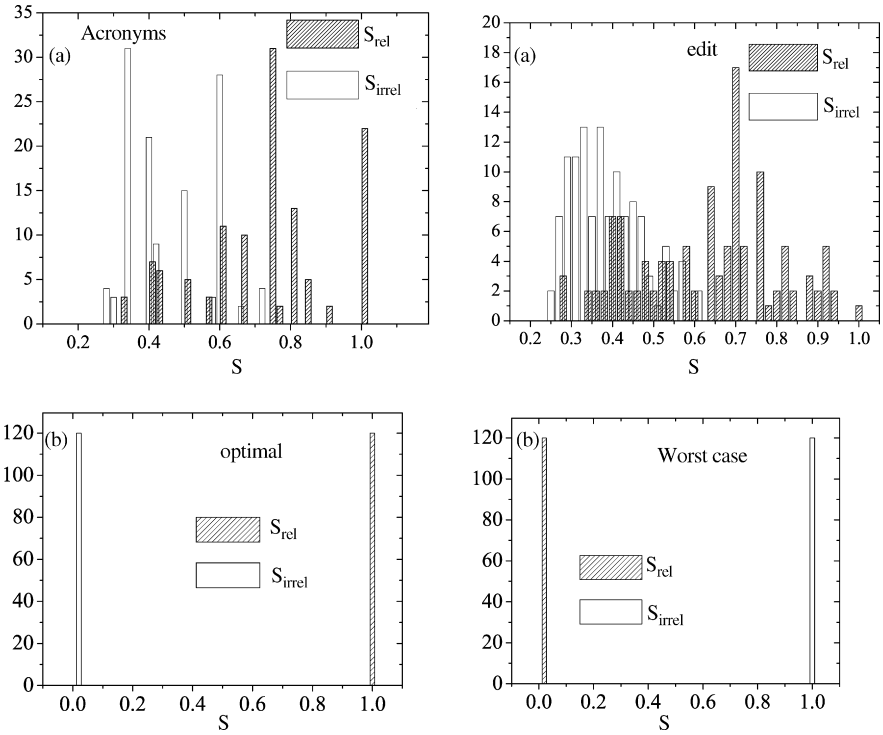


Fig. 7. The x axis represents the threshold values and the y axis represents the number of occurrences, i.e. how many queries achieved that threshold value for s_{rel} and s_{irrel}. The plots tagged (a) show examples of histograms for functions that are statistically treatable and the plots tagged (b) illustrate histograms for functions that are not statistically treatable.

likely value for t_{best} computed by our statistical method for threshold definition. In the last column of the table, we show the confidence interval built with the distribution $F(t)$ by Eqs. (10) and (11).

Table 2 shows that the absolute difference between the results of our two proposed methods for threshold estimation is at most 0.136, showing that the two approaches are in agreement. It is worth pointing out that the better the similarity function, the more in agreement the two methods are. Furthermore, in all cases the values calculated by both methods are within the theoretical confidence interval.

Table 2 also shows the results for discernability. According to them, the best real function for the data set analyzed was Jaro–Winkler and the worst was TFIDF. This can be confirmed by observing the plots in Fig. 6, which show the distribution of s_{rel} and s_{irrel}. Indeed, the best separation between relevant and irrelevant data items was achieved by Jaro–Winkler, whilst with TFIDF these items are often shuffled and/or too close together in the ranking. The plots also show the behavior of the (artificial) Optimal function, which would achieve the highest marks according to the discernability function. The behavior of Jaro, which achieved the second best result, is also plotted in Fig. 6. It is worth pointing out that this ranking is for the data set used in the experiment. For a different data set, the ranking would most probably differ.

The similarity functions in Table 2 that have the symbol * are functions for which the statistical approach is not applicable due to the nature of the data, i.e., there is no variability in the measures of similarity using this function. By no variability we mean that the values for s_{rel} and/or the values for s_{irrel} are constant for most queries. In other words, the standard deviations for s_{rel} and $s_{irrel}(\sigma(s^L_{rel})$ and $\sigma(s^L_{irrel}))$ are close to zero. Nevertheless, it is still possible to find an optimal threshold using the BestThresh algorithm.

In Fig. 7 we present plots of dispersion for the similarity values using two types of functions. Type (a) represent functions that are statistically treatable (or in which there is a reasonable variability in the data); and type (b) represent functions which are not statistically treatable.

5. Summary and conclusions

The contributions of this paper are two-fold. Our goal was to propose a method for measuring the quality of similarity functions in separating relevant from irrelevant data items returned in response to a query. In order to achieve this goal, we made a second contribution which was the development of techniques for the estimation of optimal threshold values. Such techniques can be applied not only in the evaluation of similarity functions but also in standard IR experiments to assess the quality of different ranking algorithms.

Several experiments were carried out in order to evaluate our proposed approaches. Initially, we performed experiments to test the threshold definition methods. The results show that both techniques produce similar values, validating one another.

In this paper, we used human intervention to identify relevant and irrelevant data items. However, it is worth pointing out that this sampling process could be automated through the use of clustering algorithms, as done in our previous work (Stasiu, Heuser, & da Silva, 2005). In this case, all the elements of a given cluster are considered as representing the same real world object. Thus, the relevant results for a query are the elements from the same cluster as the query. The sampling (or clustering) phase can be seen as a type of training. After this process, new queries should produce better results as a consequence of the use of a more suitable threshold.

We used the output produced by our threshold definition methods to design a "grade", which we called discernability, to measure the quality of similarity functions. The discernability takes into consideration the separation and the distance between relevant and irrelevant items. Those two aspects may be weighted differently according to their importance in the data set being analyzed. Finally, we performed experiments to assess the quality of eight similarity functions according to the discernability function. The results show that, for the data set considered, the best function was Jaro-Winkler and the worst was TFIDF.

Acknowledgments

We would like to thank the anonymous referees for the helpful suggestions.

This work was partially financed by the following projects: SisTol-CNPq, CAPES-GRICES (finished), PROBRAL-CAPES, Gerindo (CNPq/CTInfo55.2087/2002-5), XMLBroker (CNPq/Universal473310/2004-0), Rec-Semântica (FAPERGSCNPq/PRONEX-2004), Digitex (CNPq/CT-Info550845/2005-4), a PhD Scholarship from CAPES, and CAPES-PRODOC.

References

Bilenko, M., Mooney, R., Cohen, W., Ravikumar, P., & Fienberg, S. (2003). Adaptive name matching in information integration. *IEEE Intelligent Systems*, *18* (5), 16–23.

Cohen, W., Ravikumar, P., & Fienberg, S. (2003). A comparison of string distance metrics for name-matching tasks. In *Proceedings of the IIWeb* (pp. 73–78).

Dorneles, C. F., Heuser, C. A., Lima, A. E. N., da Silva, A. S., & de Moura, E. S. (2004). Measuring similarity between collection of values. In *WIDM '04: Proceedings of the sixth annual ACM international workshop on Web information and data management* (pp. 56–63). New York, NY, USA: ACM Press.

Guth, G. J. (1976). Surname spellings and computerized record linkage. *Historical Methods Newsletter*, *10* (1), 10–19.

Hall, P. A. V., & Dowling, G. F. (1980). Approximate string matching. *ACM Computing Surveys*, *12* (4), 381–402.

Jaccard, P. (1912). The distribution of flora in the alpine zone. *New Phytologist*, *11* (2), 37–50.

Jaro, M. (1989). Advances in record linking methodology as applied to the 1985 census of Tampa Florida. *Journal of the American Statistical Society*, *64*, 1183–1210.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, *10* (8), 707–710.

Navarro, G., Baeza-Yates, R., Sutinen, E., & Tarhio, J. (2001). Indexing methods for approximate string matching. *IEEE Data Engineering Bulletin*, *24* (4), 19–27.

Salton, G. (1989). *Automatic text processing: the transformation, analysis, and retrieval of information by computer*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc.

Salton, G., & McGill, M. J. (1983). *Introduction to modern information retrieval*. New York, NY, USA: McGraw-Hill.

Spiegel, M. R. (1992). *Theory and problems of probability and statistics*. McGraw-Hill.

Stasiu, R.K., Heuser, C.A., & da Silva, R. (2005). Estimating recall and precision for vague queries in databases. In *CAISE05: Proceedings of the 17th conference on advanced information systems engineering* (pp. 187–200). Springer Verlag, Porto, Portugal, June 13–17, 20, Lecture Notes in Computer Science.

Weisstein, E. W. (2004). *Bivariate normal distribution. From MathWorld—A Wolfram Web Resource*. Last modification: URL http://mathworld.wolfram.com/BivariateNormalDistribution.html.

Winkler, W. (1999). The state of record linkage and current research problems. In *Statistics of Income Division, Internal Revenue Service Publication R99/04*. URL www.census.gov/srd/www/byname.html.

Available online at www.sciencedirect.com
ScienceDirect
Journal of Informetrics 1 (2007) 26–34
Journal of INFORMETRICS An International Journal
www.elsevier.com/locate/joi

Some measures for comparing citation databases

Judit Bar-Ilan [a,*], Mark Levene [b], Ayelet Lin [a]

[a] *Department of Information Science, Bar-Ilan University, Ramat Gan 52900, Israel*
[b] *School of Computer Science and Information Systems, Birkbeck University of London, Malet Street, London WC1E 7HX, UK*

Received 29 June 2006; received in revised form 3 August 2006; accepted 3 August 2006

Abstract

Citation analysis was traditionally based on data from the ISI Citation indexes. Now with the appearance of Scopus, and with the free citation tool Google Scholar methods and measures are need for comparing these tools. In this paper we propose a set of measures for computing the similarity between rankings induced by ordering the retrieved publications in decreasing order of the number of citations as reported by the specific tools. The applicability of these measures is demonstrated and the results show high similarities between the rankings of the ISI Web of Science and Scopus and lower similarities between Google Scholar and the other tools.

Keywords: Similarity measures; Rankings; Citation databases

1. Introduction

Citation analysis is a major subfield of informetrics. Until recently the only comprehensive tool for carrying out empirical research in this area was the ISI Citation Indexes (see for example White's (2001) discussion on CAMEOs). This situation has changed, at first in individual disciplines (like CiteSeer in computer science), and now with the introduction of Elsevier's Scopus and Google Scholar.

Citation data is heavily influenced by the coverage of the specific database, since it can take into account only citations from items indexed by it. The three major tools: Web of Science (the Web version of the ISI Citation Indexes), Scopus and Google Scholar were compared and reviewed in several publications from different aspects (for example: Bauer & Bakkalbasi, 2005; Deis & Goodman, 2005; Jacso, 2005a, 2005b; Noruzi, 2005; Bar-Ilan, 2006). CiteSeer and SCISearch (a different interface of the ISI Science Citation Index) were compared by Goodrum, McCain, Lawrence, and Giles (2001). The above-mentioned studies provided numbers and descriptive statistics as a means for comparing between the different tools.

With the existence of multiple citation databases it becomes necessary to compare them systematically both from the scientometric and the informetric points of view. Descriptive statistics and specific examples are not sufficient for systematic comparison of the different citation databases. In this paper we introduce a set of measures for comparing the different citation databases. The measures compute the similarities between the rankings induced by the number of citations a publication receives in the specific database (i.e. the most cited item is ranked number 1, the second most

* Corresponding author. Tel.: +972 523667326; fax: +972 3 5353937.
E-mail addresses: barilaj@mail.biu.ac.il (J. Bar-Ilan), M.Levene@dcs.bbk.ac.uk (M. Levene), lineyal@netvision.net.il (A. Lin).

1751-1577/$ – see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.joi.2006.08.001

cited is ranked number 2, etc.). The use of these measures and statistical analysis of the results is demonstrated on a subset of the highly cited Israeli researchers, as defined in ISI's Highly Cited database (ISI HighlyCited.com, 2002) supplemented by the three recent Israeli Nobel prize winners.

The measures are defined in Section 2, the data collection and empirical settings appear in Section 3. In Section 4 the results are displayed and analyzed, and Section 5 concludes the paper.

2. The measures

The rankings were compared using four basic measures that complement each other. In this section the measures are defined. Each of the measures is defined for a pair of databases (A and B), where A and B can WoS (Web of Science), Scopus or Google Scholar. The measures introduced here were applied to comparing rankings of search engine rankings (Bar-Ilan, Mat-Hassan, & Levene, 2006; Bar-Ilan, Levene, & Mat-Hassan, 2006; Bar-Ilan, Keenoy, Yaari, & Levene, submitted for publication).

2.1. Overlap and footrule

Overlap (O) is defined as follows:

$$O = \frac{|\mathrm{PUBL_A} \cap \mathrm{PUBL_B}|}{|\mathrm{PUBL_A} \cup \mathrm{PUBL_B}|}$$

where PUBL_X is the set of publications retrieved from database X. The measure O does not take into account the rankings, it only measures the proportion of the publications retrieved from both databases out of the total number of publications retrieved by either of them.

Footrule, F, is the normalized Spearman footrule. Spearman's footrule (Diaconis & Graham, 1977; Dwork, Kumar, Naor, & Sivakumar, 2001) can be computed for two permutations, and thus it can be applied only for the publications that are ranked in both databases. Each such publication is given its relative rank in the set of publications retrieved from both databases. Suppose for the moment that there are no ties in the rankings (i.e. no two publications receive exactly the same number of items). This is an unrealistic assumption and we will deal with it in Section 3. The result of the re-rankings is two permutations σ_1 and σ_2 on $1 \ldots Z$ where $|Z|$ is the number of overlapping publications. After these transformations Spearman's footrule is computed as

$$Fr^{|Z|}(\sigma_1, \sigma_2) = \sum_{i=1}^{|Z|} |(\sigma_1(i) - \sigma_2(i))|$$

When the two rankings are identical on the set Z, $Fr^{|Z|}$ is zero, and its maximum value is $|Z|^2$ when $|Z|$ is even, and $(|Z|+1)(|Z|-1)$ when $|Z|$ is odd. When the result is divided by its maximum value, $Fr^{|Z|}$ will be between 0 and 1, independent of the size of the overlap. This measure is undefined for $|Z| = 0,1$. Thus we compute the *normalized Spearman's footrule*, NFr, for $|Z| > 1$

$$NFr = \frac{Fr^{(|Z|)}}{\max Fr^{(|Z)}}$$

NFr ranges between 0 and 1; it attains the value 0 when the relative ranking of the publications in the set Z is identical. Since we are interested in similarity measures, we define F as

$$F = 1 - NFr$$

The weakness of this measure is that it totally ignores the non-overlapping elements and only takes into account the relative rankings, thus for example if $|Z| = 2$, and these two publications are ranked at ranks 1 and 2 in database A, while in database B they are ranked at 9 and 10 (and the first eight publications are not ranked in database A), the value of F will be 1, just like the case where both A and B rank these two publications at ranks 1 and 2, respectively.

2.2. Fagin measure

Spearman's footrule is a very useful measure to compare the ordering in two permutations (Diaconis & Graham, 1977; Dwork et al., 2001). However when comparing two sets of ranked results the underlying sets are often not identical. Fagin, Kumar, and Sivakumar (2003) extended Spearman's footrule in such a way that the measure does not require that both rankers rank exactly the same set of items. They developed the measure for comparing search engine rankings, but here we modify the description to fit the current setting. Suppose that exactly k publications were retrieved from both databases (not necessarily the same publications). The number of citations each publication receives induces a natural ranking on these items. Suppose for the moment that there are no ties in the rankings (i.e. no two publications receive exactly the same number of items). This is an unrealistic assumption and we will deal with it in Section 3. Each publication that was retrieved from A, but not from B is artificially assigned rank $k+1$ (similarly for the publications retrieved from B and not from A). The rationale for this artificial rank is that if A indexes the specific item its rank would be $k+1$ or more (note that if it is not indexed by A it would not be ranked at all).

Let Z be the set of publications retrieved by both databases, S the set of publications retrieved only from A and T the set of publications retrieved only from B, σ_1 the ranking of the publications retrieved from A and σ_2 the ranking of the publications retrieved from B then

$$F^{(k+1)}(\sigma_1, \sigma_2) = \sum_{i \in Z} |\sigma_1(i) - \sigma_2(i)| + \sum_{i \in S} ((k+1) - \sigma_1(i)) + \sum_{i \in T} ((k+1) - \sigma_2(i))$$

This measure has to be normalized so that when the two rankings are on identical sets in identical order the measure equals 1, and when there is no overlap between the sets, the measure is 0. Thus

$$G^{(k+1)} = 1 - \frac{F^{(k+1)}}{\max F^{(k+1)}}$$

where $\max F^{(k+1)} = k(k+1)$.

In case the number of retrieved elements from both databases is not identical, we introduce the following modification of the measure: suppose k_1 publications were retrieved from database A and k_2 from database B, then

$$F^{(k_1 k_2)}(\sigma_1, \sigma_2) = \sum_{i \in Z} |\sigma_1(i) - \sigma_2(i)| + \sum_{i \in S} ((k_2+1) - \sigma_1(i)) + \sum_{i \in T} ((k_1+1) - \sigma_2(i))$$

and

$$G^{(k_1,k_2)} = 1 - \frac{F^{(k_1,k_2)}}{\max F^{(k_1,k_2)}}$$

where

$$\max F^{(k_1,k_2)} = \frac{k_1(k_1+1)}{2} + \frac{k_2(k_2+1)}{2}$$

This measure unlike the footrule, takes into account the non-overlapping publications as well, but in our opinion it gives too much weight to these elements. Suppose that, for two lists of ten ranked results, $|Z| = 5$, and A ranks z_1 at position 1, z_2 at position 2 ... and z_5 at position 5; B ranks z_1 at position 5, z_2 at position 4 ... and z_5 at position 1. In this case $G^{(10,10)}$ equals 0.618. Now if B ranks the items in Z exactly like A, $G^{(10,10)}$ increases only slightly to 0.727. The amount of change in G for a given overlap is rather small, since G is mainly determined by the size of the overlap.

2.3. Inverse rank measure

Our last measure attempts to correct this problem, by giving more weight to identical or near identical rankings among the top ranking publications. This measure tries to capture the intuition that identical or near identical rankings among the top publications indicate greater similarity between the rankings induced by the databases. First, let

$$N^{(k_1,k_2)}(\sigma_1\sigma_2) = \sum_{i \in Z} \left| \frac{1}{\sigma_1(i)} - \frac{1}{\sigma_2(i)} \right| + \sum_{i \in S} \left| \frac{1}{\sigma_1(i)} - \frac{1}{(k_2+1)} \right| + \sum_{i \in T} \left| \frac{1}{\sigma_2(i)} - \frac{1}{(k_1+1)} \right|$$

where k_1, k_2, S, T, Z and σ_i are as before. This measure has to be normalized as well, thus

$$M^{(k_1,k_2)} = 1 - \frac{N^{(k_1,k_2)}}{\max N^{(k_1,k_2)}}$$

where

$$\max N^{(k_1,k_2)} = \sum_{i=1}^{k_1} \left(\frac{1}{i} - \frac{1}{k_2+1} \right) + \sum_{i=1}^{k_2} \left(\frac{1}{i} - \frac{1}{k_1+1} \right)$$

Considering the same two cases as before (five overlapping elements, opposite versus identical rankings), the M values will be 0.386 and 0.905, respectively, emphasizing the importance of similarity in rankings in the top positions. Now suppose that $|Z| = 5$ as before, but the overlapping elements are ranked 6, 7, 8, 9 and 10 by both A and B. In this case $G^{(10,10)}$ is 0.182 (compared with 0.727 when the overlapping elements were identically ranked in the top positions) while $M^{(10,10)}$ is 0.149 (compared with 0.905 when the overlapping elements were identically ranked in the top positions)—showing the larger weight given by the M measure to overlapping elements in the top positions.

3. Data collection

To demonstrate the feasibility of the measures, we applied them to the publications of the highly cited Israeli scientists (ISI HighlyCited.com, 2002), as defined by the ISI (ISI HighlyCited.com). This list is based on citations to items indexed by ISI and were published between 1981 and 1999. The list is comprised of 44 names. Rather interestingly it does not include the three Israeli Nobel prize winners in the last two years (Robert Aumann, Aaron Ciechanover and Avram Hershko). These three names were added to the list. We had disambiguation problems with a few of the names, and as a result we excluded eight names.

Scopus only provides full citation data of items from 1996 and onwards. In order to have a "fair" comparison, only publications from 1996 and onwards were considered. Note that this is a different period from the period for which the researcher was included in the list of highly cited authors, thus it may well be the case that during the period under consideration the researcher will have few or no publications.

We only considered highly cited papers of the highly cited researchers, and thus only items with 20 or more citations in the specific database were retrieved. There were two reasons for this decision: (1) the data had to be carefully cleansed (especially from Google Scholar) and by considering only the most highly cited items, we were able to carry out the cleansing in reasonable time and (2) as noted in the previous section we had to find a solution for "ties", i.e. two publications that received the same number of citations in the specific database. Among the more highly cited items there were fewer occurrences of ties. Note that as a result of the decision to retrieve publications with twenty or more citations only, we do not have the information whether an item retrieved from database A, and not retrieved from database B is indexed by B (but received less than 20 citations from sources indexed by B) or is not indexed at all by B. This is the usual assumption when applying the measures discussed in Section 2.

In the final list for analysis we only included scientists whose publications appeared in all three databases and had at least three items published from 1996 onwards with 20 or more citations. Thus we had to exclude 16 additional names: two researchers had no highly cited publications in any of the three databases, one had 60 highly cited papers in WOS, two in Google Scholar and none in Scopus (a physicist), 12 (computer scientists and/or mathematicians, and one pharmacologist) had a considerable number of highly cited publications indexed by Google Scholar, but less than three by either WOS or Scopus (usually both). These differences are due to the fact that Google indexes books, proceedings and technical reports as well, but WOS excludes these types of publications almost entirely and Scopus indexes them only in a limited fashion. There was only a single case where the scientist (a hydrologist) was excluded because of the lack of enough highly cited items indexed by Google Scholar. The final list was comprised of 22 scientists.

Table 1 lists these scientists together with the number of items with more than 20 citations from each database and the total number of citations received by these items. The searches were carried out during the second half of January 2006. There is no clear "winner", but it seems that Google Scholar retrieves more items and citations in computer science and less in chemistry. The differences between the Web of Science and Scopus are not as significant. An interesting case is Ehud Duchovni—he is a high energy physicist, a member of the Opal and Atlas groups conducting experiments at

Table 1
The list of scientists examined, the number of highly cited publications and the total number of citations these publications received in each citation database

Scientist	Affiliation	Discipline	WOS		Scopus		Google Scholar	
			Items	Total citations	Items	Total citations	Items	Total citations
Alon, Noga	Tel Aviv U.	Mathematics, computer science	8	220	9	274	30	1438
Aurbach, Doron	Bar Ilan U.	Materials science	40	2073	40	2067	18	562
Chet, Ilan	Weizmann Inst.	Plant & animal science	17	552	17	593	16	567
Ciechanover, Aaron	Technion	Molecular biology & genetics	38	6194	41	6309	35	5202
Cohen, Irun R.	Weizmann Inst.	Immunology	37	2210	40	2357	34	1910
Dekel, Avishai	Hebrew U.	Space sciences	27	1618	18	1162	14	1220
Duchovni, Ehud	Weizmann Inst.	Physics	63	2352	29	1038	47	1891
Geiger, Benjamin	Weizmann Inst.	Molecular biology & genetics	44	3836	45	3853	42	3282
Goldreich, Oded	Weizmann Inst.	Computer science	8	326	5	246	39	2957
Harel, David	Weizmann Inst.	Computer science	3	112	4	264	23	3090
Hershko, Avram	Technion	Molecular biology & genetics	21	3743	21	3705	20	2888
Jortner, Joshua	Tel Aviv U.	Chemistry	28	1760	25	1436	15	621
Kanner, Joseph	Agricultural Research Organization	Agricultural sciences	4	195	4	201	3	88
Kerem, Batsheva	Hebrew U.	Molecular biology & genetics	18	969	17	943	14	673
Mechoulam, Raphael	Hebrew U.	Pharmacology	30	2110	33	2393	30	1543
Oren, Moshe	Weizmann Inst.	Molecular biology & genetics	66	7021	65	7227	61	6156
Piran, Tsvi	Hebrew U.	Space sciences	38	3016	28	1807	30	2943
Procaccia, Itamar	Weizmann Inst.	Physics	12	439	13	510	13	476
Shamai, Shlomo	Technion	Computer science	12	658	14	1083	22	1961
Sharir, Micha	Tel Aviv U.	Engineering, computer science	4	114	7	175	20	782
Sklan, David	Hebrew U.	Agricultural sciences	11	311	13	346	4	102
Turkel, Eli	Tel Aviv U.	Mathematics	3	90	4	124	4	138

CERN in Geneva. There are more than one hundred members in these groups, and all of them coauthor each publication. The Web of Science indexes all the authors, while Scopus does not. Another Israeli member of these groups is Giora Mikenberg (also in the list of highly cited researchers). The list of highly cited items he authored is highly similar to the list of Ehud Duchovni on Web of Science, but there were no items retrieved from Scopus—probably due to the fact that his name is further down on the list (M vs. D!) and therefore the Opal and Atlas publications were not attributed to him on Scopus.

The measures described in Section 2 can only be applied to ranked lists without ties. When the ranking is induced by the number of citations received there are often ties. Ties were resolved in the following way: suppose items x and y were retrieved by database A and have exactly the same citation count.

Case 1. *x* and *y* were also retrieved by B

a) *x* received more citations than *y* in database B. In this case $\text{rank}_A(x)<\text{rank}_A(y)$ (recall that the lower rank numbers correspond to higher citation counts).
b) *y* received more citations than *x* in database B. In this case $\text{rank}_A(y)<\text{rank}_A(x)$ (recall that the lower ranks correspond to higher citation counts).
c) *x* and *y* were tied also in B. In this case the decision is arbitrary, but consistent in both lists, i.e. either $\text{rank}_A(x)<\text{rank}_A(y)$ and $\text{rank}_B(x)<\text{rank}_B(y)$ or $\text{rank}_A(y)<\text{rank}_A(x)$ and $\text{rank}_B(y)<\text{rank}_B(x)$.

Case 2. Only one of the items, say x, is retrieved by B, then $\text{rank}_A(x)<\text{rank}_A(y)$.

Case 3. Neither x nor y are retrieved by B, then the decision is arbitrary.

This tie-resolution algorithm can be easily extended to the case where more than two items are tied. Note that the tie-resolution is dependent on the database B, thus different rankings may result when comparing the results of A to B or to C.

4. Results

The four measures introduced in Section 2 were computed for each researcher and for each pair of databases. The results are displayed in Table 2. Note that values above 0.7 indicate high similarity; this threshold was set arbitrarily and is similar to the accepted threshold for "high correlation".

We see that there is complete agreement between Scopus and the Web of Science on the ranked lists of Avram Hershko and Joseph Kanner. In Kenner's case the list is only four items long, but in Hershko's case we are comparing lists of length 21. When looking at the number of citations on a per item basis, we see that the Web of Science recorded slightly more citations than Scopus, for example for the top ranked item, the Web of Science reported 1919 citations, whereas Scopus reported 1909 citations. There was almost total agreement on Eli Turkel's list as well, the only difference was that WoS listed only three publications with more than 20 citations, and Scopus listed four. Checking WoS we observed that the fourth item on Scopus is indeed number four on the WoS list as well, but was not retrieved because it received only 18 citations.

There was much less agreement between Google Scholar and the other two databases. One of the most striking results is Oded Goldreich. Scopus listed 5 items and Google Scholar 39 items, with only two overlapping items (publications that were retrieved by both database) and these two items appeared in opposite order in the two lists

Table 2
Similarity values for the ranked lists from Web of Science, Scopus and Google Scholar

Researcher	Web of Science – Scopus				Web of Science – Google Scholar				Scopus – Google Scholar			
	O	*F*	*G*	*M*	*O*	*F*	*G*	*M*	*O*	*F*	*G*	*M*
Alon	0.545	0.778	0.725	0.584	0.276	0.750	0.457	0.468	0.267	0.750	0.466	0.703
Aurbach	0.951	0.942	0.972	0.982	0.415	0.639	0.652	0.552	0.415	0.694	0.663	0.618
Chet	0.889	0.922	0.961	0.884	0.833	0.839	0.872	0.842	0.737	0.796	0.889	0.816
Ciechanover	0.927	0.981	0.969	0.989	0.775	0.921	0.928	0.916	0.762	0.926	0.941	0.922
Cohen	0.925	0.942	0.964	0.963	0.821	0.840	0.873	0.922	0.762	0.855	0.874	0.920
Dekel	0.667	0.938	0.767	0.828	0.323	0.760	0.468	0.574	0.231	1	0.412	0.647
Duchovni	0.394	0.728	0.535	0.377	0.528	0.825	0.736	0.589	0.357	0.700	0.480	0.304
Geiger	0.978	0.973	0.987	0.980	0.870	0.918	0.926	0.919	0.891	0.907	0.920	0.912
Goldreich	0.444	1.000	0.792	0.911	0.119	0.833	0.222	0.157	0.073	0	0.146	0.112
Harel	0.750	0.500	0.600	0.353	0.130	1	0.237	0.134	0.174	0.750	0.322	0.280
Hershko	1	1	1	1	0.952	0.880	0.941	0.949	0.952	0.880	0.932	0.946
Jortner	0.893	0.878	0.891	0.875	0.448	0.643	0.686	0.701	0.444	0.528	0.639	0.604
Kanner	1	1	1	1	0.750	1	0.867	0.928	0.750	1	0.867	0.928
Kerem	0.944	0.972	0.895	0.791	0.778	0.918	0.787	0.774	0.824	0.939	0.855	0.897
Mechoulam	0.853	0.962	0.961	0.966	0.781	0.821	0.892	0.904	0.765	0.799	0.898	0.889
Oren	0.985	0.961	0.961	0.913	0.866	0.904	0.911	0.885	0.881	0.909	0.925	0.953
Piran	0.585	0.910	0.710	0.716	0.457	0.827	0.639	0.691	0.400	0.797	0.616	0.792
Procaccia	0.923	0.944	0.964	0.978	0.786	0.800	0.887	0.853	0.733	0.800	0.881	0.853
Shamai	0.857	0.750	0.884	0.863	0.545	0.639	0.754	0.790	0.636	0.878	0.851	0.921
Sharir	0.571	0.500	0.800	0.519	0.091	1	0.216	0.447	0.125	0.500	0.271	0.304
Sklan	0.500	0.875	0.787	0.871	0.250	0.500	0.507	0.663	0.308	0.750	0.565	0.700
Turkel	0.750	1	1	1	0.400	1	0.533	0.331	0.600	1	0.600	0.377
Average	0.788	0.884	0.869	0.834	0.554	0.830	0.681	0.681	0.549	0.780	0.682	0.700
S.D.	0.198	0.147	0.136	0.200	0.278	0.134	0.242	0.249	0.276	0.220	0.247	0.262

Table 3
Results of the statistical tests for the measures O, G and M[a]

		O	G	M
WoS-Scopus	Mean	0.788	0.869	0.834
	S.D.	0.198	0.136	0.200
WoS-GS	Mean	0.554	0.681	0.681
	S.D.	0.278	0.242	0.249
Scopus-GS	Mean	0.549	0.682	0.700
	S.D.	0.296	0.247	0.262
F		38.266***	21.674***	8.433**
WoS-Scopus vs. WoS-GS		***	***	**
WoS-Scopus vs. Scopus-GS		***	***	*
WoS-GS vs. Scopus-GS		–	–	–

(*), (**) and (***) indicate the strength of the significance, (–) indicates that no differences were found. Levels of significance: $^{*}p<.05$, $^{**}p<.01$, $^{***}p<.001$.

[a] WoS-Scopus vs. WoS-GS checks whether the significant F-value was caused by differences in the WoS-Scopus vs. WoS-GS measures.

(that is why $F=0$). The G and M values are very low because of the extremely small overlap and the disagreement in the ordering of the overlapping elements. Note that the items listed by Google Scholar were checked against Oded Goldreich's list of publication and/or the publisher's site, non-existing publications were removed and publications listed more than once were collated. We observe a similar pattern when comparing David Harel's list in WoS versus Google Scholar. In this case there were three overlapping elements (all the elements listed by WoS), but in his case there was total agreement on the relative ranking, i.e. the items that were ranked 1, 2 and 3 respectively on WoS, were ranked 3, 5 and 6 on Google Scholar. Thus in this case $F=1$, but the G and M values are low, because Google Scholar ranked 23 publications versus 3 by WoS and there was no agreement on the top ranked items of Google Scholar (these were not listed by WoS). Note that the most cited item according to Google Scholar is a book and WoS does not index books.

From looking at the averages in Table 2, it seems that the WoS and Scopus rankings are rather similar, whereas the Google Scholar ranking is considerably different. However, the standard deviations are considerable. Thus we decided to run some statistical tests. We ran the repeated measure ANOVA test for each query, for each database and for each measure, to compare the rankings of the databases for the 22 researchers. This test measures the variability of the database rankings (i.e. whether the similarity between the rankings of database A and B is significantly different from the similarity between the rankings of database A and C); see for example (Grimm & Yarnold, 2005) If the results of this F-test (ANOVA) is significant then it means that the differences between the three pairs WoS-Scopus, WoS-GS (GS stands for Google Scholar) and Scopus-GS cannot be explained by random errors, but there are consistent differences. If the F-value is significant, one can run additional tests; in our case the appropriate tests were Bonferroni-adjusted post-hoc tests to determine the differences between which two pairs are responsible for the significance of the F-test. There are cases where the differences between more than two pairs are significant. The tests were run on all four measures, however the footrule (F) did not fulfill the test assumptions (sphericity), and thus here we report only the significance of the statistical tests for three measures, O, G and M (see Table 3).

The results indicate that the significant differences for the set of researchers tested for this study were caused by the considerable differences in the rankings of Google Scholar as compared to either the Web of Science or Scopus.

Google Scholar (and to some extent Scopus as well) indexes proceedings and books as well, while WoS indexes mainly journal papers. For three researchers: Alon, Goldreich and Harel the number of items indexed by Google Scholar was considerably higher than the number of items indexed by the other two databases. For these researchers we compared the rankings induced by the three databases, when considering journal papers only. The results appear in Tables 4 and 5.

As can be seen in Table 5, the differences remain considerable even when only journal publications are taken into account.

Table 4
Number of journal papers indexed by each database

Scientist	Affiliation	Discipline	WOS		Scopus		Google Scholar	
			Items	Total citations	Items	Total citations	Items	Total citations
Alon, Noga	Tel Aviv U.	Mathematics, computer science	8	220	8	243	20	210
Goldreich, Oded	Bar Ilan U.	Computer science	8	302	4	220	7	776
Harel, David	Weizmann Inst.	Computer science	3	112	4	264	9	1613
Sharir, Micha	Technion	Engineering, computer science	4	114	7	175	11	457

Table 5
Similarity values for the ranked lists from Web of Science, Scopus and Google Scholar when considering journal papers only

Researcher	Web of Science – Scopus				Web of Science – Google Scholar				Scopus – Google Scholar			
	O	*F*	*G*	*M*	*O*	*F*	*G*	*M*	*O*	*F*	*G*	*M*
Alon	0.600	0.778	0.778	0.615	0.400	0.750	0.615	0.545	0.400	0.750	0.641	0.803
Goldreich	0.571	1.000	0.905	0.965	0.400	0.750	0.556	0.370	0.222	0.000	0.400	0.287
Harel	0.750	0.500	0.600	0.353	0.333	1.000	0.533	0.358	0.444	0.750	0.760	0.869
Sharir	0.571	0.500	0.800	0.519	0.154	1.000	0.377	0.601	0.200	0.500	0.422	0.362

5. Conclusions

In this paper we introduced a set of measures for comparing rankings of different citation databases induced by the number of citations the tested publications receive in each database.

The results indicate that Scopus and the Web of Science are comparable in terms of the rankings induced. Note that the measures were computed only for a small set of cases, in order to be able to generalize the results, larger-scale, discipline-specific tests should be carried out.

Some of the differences are caused by the differing indexing strategies of the databases. Google Scholar does not have a clear policy, but unlike WoS it indexes books and proceedings as well. These types of publications are often cited more than journal papers, especially in computer science. Had we compared the rankings only on journal papers, the similarity measures.

References

Bar-Ilan, J. (2006). H-index for Price medalists revisited. *ISSI Newsletter*, *2*(1), 3–5.

Bar-Ilan, J., Levene, M., & Mat-Hassan, M. (2006). Methods for evaluating dynamic changes in search engine rankings – A case study. *Journal of Documentation*, *62*(6).

Bar-Ilan, J., Keenoy, K., Yaari, E., & Levene, M. (submitted for publication). User rankings of search engine results.

Bar-Ilan, J., Mat-Hassan, M., & Levene, M. (2006). Methods for comparing search engine results. *Computer Networks*, *50*(10), 1448–1463.

Bauer, K., & Bakkalbasi, N. (2005). An examination of citation counts in a new scholarly communication environment. *D-Lib Magazine*, *11*(9). Retrieved June 16, 2006, from http://www.dlib.org/dlib/september05/bauer/09bauer.html.

Deis, L. F., & Goodman, D. (2005). Web of Science (2004 version) and Scopus. *The Charleston Advisor*, *6*(3), 5–21. Retrieved June 16, 2006, from http://www.charlestonco.com/comp.cfm?id=43.

Diaconis, P., & Graham, R. L. (1977). Spearman's footrule as a measure of disarray. *Journal of the Royal Statistical Society, Series B (Methodological)*, *39*, 262–268.

Dwork, C., Kumar, R., Naor, M., & Sivakumar, D. (2001). Rank aggregation methods for the Web. In *Proceedings of the 10th World Wide Web Conference* (pp. 613–622).

Fagin, R., Kumar, R., & Sivakumar, D. (2003). Comparing top k lists. *SIAM Journal on Discrete Mathematics*, *17*(1), 134–160.

Goodrum, A. A., McCain, K. W., Lawrence, S., & Giles, L. C. (2001). Scholarly publishing in the Internet age: A citation analysis of computer science literature. *Information Processing and Management*, *37*(6), 661–675.

Grimm, L. G., & Yarnold, P. R. (2005). *Reading and understanding multivariate statistics*. Washington, DC: American Psychological Association.

ISI HighlyCited.com (2002). ISI Highly Cited Researchers – Country: Israel. Retrieved June 16, 2006, from http://hcr3.isiknowledge.com/browse_author.pl?page=0&link1=Browse&valueCategory=0&valueCountry=81&submitCountry.x=18&submitCountry.y=6.

ISI HighlyCited.com. About ISI HighlyCited.com. Retrieved June 16, 2006, from http://hcr3.isiknowledge.com/popup.cgi?name=hccom.

Jacso, P. (2005). As we may search—Comparison of major features of Web of Science, Scopus and Google Scholar citation-based and citation-enhanced databases. *Current Science*, *89*(9), 1537–1547.

Jacso, P. (2005b). Comparison and analysis of the citedness scores in Web of Science and Google Scholar. In Proceeding of Digital Libraries: Implementing Strategies and Sharing Experiences, *Lecture Notes in Computer Science*, 3815, 360–369.

Noruzi, A. (2005). Google Scholar: The new generation of citation indexes. *LIBRI*, *55*(4), 170–180.

White, H. D. (2001). Author-centered bibliometrics through CAMEOs: Characterizations automatically made and edited online. *Scientometrics*, *51*(3), 607–637.

Available online at www.sciencedirect.com
ScienceDirect
Journal of Informetrics 1 (2007) 2–7
Journal of INFORMETRICS An International Journal
www.elsevier.com/locate/joi

The influence of missing publications on the Hirsch index

Ronald Rousseau [a,b,c]

[a] *KHBO (Association K.U.Leuven), Department of Industrial Sciences and Technology, Zeedijk 101, B-8400 Oostende, Belgium*
[b] *University of Antwerp (UA), IBW, Universiteitsplein 1, B-2610 Wilrijk, Belgium*
[c] *Hasselt University (UHasselt), Agoralaan, Building D, B-3590 Diepenbeek, Belgium*

Received 2 May 2006; received in revised form 30 May 2006; accepted 30 May 2006

Abstract

We show that usually the influence on the Hirsch index of missing highly cited articles is much smaller than the number of missing articles. This statement is shown by a combinatorial argument. We further show, by using a continuous power law model, that the influence of missing articles is largest when the total number of publications is small, and non-existing when the number of publications is very large. The same conclusion can be drawn for missing citations. Hence, the *h*-index is resilient to missing articles and to missing citations.

Keywords: Hirsch index; Influence of missing data; Power law model; Robustness

1. Introduction

Recently the Hirsch index, in short: *h*-index, has attracted a lot of attention in the scientific community (Bar-Ilan, 2006; Egghe, in press; Glänzel, 2006; Liang, 2006). This index, introduced by Hirsch (2005) is calculated as follows. Consider the list of publications co-authored by scientist S, ranked according to the number of citations each of these has received over a given period. Then scientist S' *h*-index is *h* if it is the largest natural number such that the first *h* publications received each at least *h* citations. Clearly, this definition can also be applied to some other source-item pairs, besides a scientist's publications and citations (Braun et al., 2005; Egghe & Rousseau, 2006; Rousseau, 2006).

In most applications citations have been taken into account only if the corresponding articles have been published in a journal covered by the Web of Knowledge (Thomson Scientific). Yet, it is also possible to collect citations from the Web via Google Scholar (Bar-Ilan, 2006), or from a local database such as the CSCD in China (Liu Zeyuan, personal communication). Expressed in a conglomerate framework this means that the used pool is essential (Rousseau, 2005). It is indeed quite feasible that a scientist's most cited works are published in conference proceedings, free web-journals or, generally, in sources not covered by the Web of Knowledge. What then is the influence of these highly cited articles on a scientist's *h*-index?

E-mail address: ronald.rousseau@khbo.be.

1751-1577/$ – see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.joi.2006.05.001

2. A simple discrete model

It is assumed that the number of missing articles, denoted as m, contains s highly cited ones, this is: articles above the level of the h-index. Secondly, it is assumed that in the initial situation the article at rank h receives exactly h citations. Finally, it is assumed that in the neighbourhood of the original h-index the difference between the numbers of citations received by consecutive articles in the ranking is a fixed number. None of these assumptions is crucial for the point we want to make, namely that the difference in ranking resulting from the missing articles is never equal to the number of missing articles, and usually much smaller. We note already that if s is zero the h-index remains the same. This happens when all missing articles receive a relatively low number of citations.

Case A. The h-index falls in a zone where there are many articles ($>s$) that received the same number of citations. In that case inclusion of the missing articles will result either in the same h-index or in an h-index that is one unit higher. Table 1 illustrates Case A.

Case B. The h-index falls in a zone (we assume that the length of that part of this zone before h is larger than or equal to s) where consecutive articles differ by exactly one citation received. In that case inclusion of the missing articles will result either in an increase of the h-index by $s/2$ (if s is even) or by $(s-1)/2$ (s is odd).

Indeed, assume that $s=2n$. Then after inclusion of s highly cited articles, the article originally at rank h is now ranked $h+2n$, and has still h citations. The article that occurred at rank $h+1$, now has rank $h+2n+1$, and has $h-1$ citations. In general, the article now at rank $h+2n+k$ receives $h-k$ citations. Hence, in order to determine the new h-index we solve: $h+2n+k=h-k$, yielding $k=-n$. So the new h-index is $h+n=h+s/2$. If s is odd, say $2n+1$, then we have to solve $h+2n+1+k=h-k$, yielding $k=-(2n+1)/2$. Because the standard Hirsch index is a natural number, also in this case the change is only equal to $n=(s-1)/2$, and not equal to s as perhaps expected. Table 2 gives an example for $s=3$.

Case C. Larger gaps

We assume that, in the original ranking, before the article at rank h there is, over the zone of interest, always a gap between the number of citations equal to G (>1). Clearly, Case B is the case $G=1$.

Table 1
An illustration of Case A for $s=3$

(a) h-index remains unchanged

Rank	Number of citations	Rank (including s new highly cited articles)	Number of citations
...	...	...	...
$h-3$	h	$h-3$	*
$h-2$	h	$h-2$	*
$h-1$	h	$h-1$	*
h	h	h	h
$h+1$	h	$h+1$	h

(b) h-index changes one unit

Rank	Number of citations	Rank (including s new highly cited articles)	Number of citations
...	...	...	...
$h-4$	$h+1$	$h-4$	*
$h-3$	$h+1$	$h-3$	*
$h-2$	$h+1$	$h-2$	*
$h-1$	h	$h-1$	$h+1$
h	h	h	$h+1$
$h+1$	h	$h+1$	$h+1$
$h+2$	h	$h+2$	h

* Exact value does not matter.

Table 2
An illustration of Case B for $s = 2n + 1 = 3$

Rank	Number of citations	Rank (including s new highly cited articles)	Number of citations
...	...	...	...
$h-4$	$h+4$	$h-4$	*
$h-3$	$h+3$	$h-3$	*
$h-2$	$h+2$	$h-2$	*
$h-1$	$h+1$	$h-1$	$h+4$
h	h	h	$h+3$
$h+1$	*	$h+1$	$h+2$
$h+2$	*	$h+2$	$h+1$

Here the new h-index is equal to $h+1$.
* Exact value does not matter.

If $s < G$ then the article formerly situated at rank h is now at rank $h+s$ and still receives h citations. The article formerly situated at rank $h-1$ moves to rank $h+s-1$ and has $h+G$ citations. Consequently, if $s \leq G$ then the new h-index is equal to $h+s-1$. Note that this means that h remains unchanged if $s = 1$. If $G < s \leq 2G$ then, one can easily see that the new h-index is $h+s-2$, and if $2G < s \leq 3G$ then the new h-index is $h+s-3$. This shows that even for relatively large gaps the influence of s missing highly cited articles is not equal to s, but smaller. Table 3 provides an example.

3. A first example: Citations follow a Zipf distribution

If citations follow a Zipf distribution this means that the number of citations of the source at rank r is equal to Z/r. In this case the h-index is found by solving the equation $h = Z/h$, hence $h = \sqrt{Z}$. Taking h equal to a natural number means that h is equal to the largest natural number smaller than or equal to $\sqrt{Z}$. This number is known as the floor function of $\sqrt{Z}$ denoted as $\lfloor \sqrt{Z} \rfloor$. In Table 4 the h-index is calculated for some values of Z, as well as the number of citations (rounded) of the sources at rank $h-1$. This allows us to assess if this situation corresponds (approximately) to Cases A, B or C. Recall that $h(1) = Z$, which corresponds to the source with the highest production (in this example: the article receiving the most citations).

Table 4 shows that for these realistically highest numbers of citations (Z-values) the gaps are rather small so that if, for example, four highly cited publications are missing, then this would result in a change of the h-index equal to one to three.

Table 3
An illustration of Case C for $G = 3$ and $s = 8$

Rank	Number of citations	Rank (including $s = 8$ new highly cited articles)	Number of citations
...	...	...	...
$h-4$	$h+12$	$h-4$	*
$h-3$	$h+9$	$h-3$	*
$h-2$	$h+6$	$h-2$	*
$h-1$	$h+3$	$h-1$	*
h	h	h	*
$h+1$	*	$h+1$	*
$h+2$	*	$h+2$	*
$h+3$	*	$h+3$	*
$h+4$	*	$h+4$	$h+12$
$h+5$	*	$h+5$	$h+9$
$h+6$	*	$h+6$	$h+6$
$h+7$		$h+7$	$h+3$
$h+8$	*	$h+8$	h

The new h-index is equal to $h+s-2 = h+6$.

Table 4
The Zipf model for the *h*-index

Z	*h*	# citations at rank $h-1$
2000	44	47
1000	31	33
500	22	24
200	14	15
100	10	11
50	7	8

4. A second example: Price awardees

Leo Egghe has recently introduced an alternative for the *h*-index (Egghe, 2006a, 2006b, 2006c). This is not the subject of this note, but we will use his tables of *h*-indices of Price medallists to study the influence of missing publications on a scientist's *h*-index. We will assume that for each of them five highly cited articles are missing and we will recalculate their *h*-index, based on the data in (Egghe, 2006c). Results are shown in Table 5.

This example shows that for this list of real *h*-indexes an additional five articles yields an increase in the *h*-index value between zero and four. On average the increase is 2.21 or less than half the number of missing publications.

5. An analytical model based on a power law

In this section we show that a power law, i.e. a Lotka model, as used in an earlier publication (Egghe & Rousseau, 2006) leads to the same conclusion as the combinatorial argument presented above.

In this earlier publication we proved that if citations (or in general: item frequencies) can be described by a negative power law with exponent $\alpha > 1$, and if the system has T sources, then the *h*-index (actually its real-valued version, because in this approach the *h*-index is not a natural number anymore) is equal to

$$h_1 = T^{1/\alpha} \tag{1}$$

Adding m missing articles (sources), and assuming that this addition does not alter the exponent α, then the *h*-index becomes:

$$h_2 = (T+m)^{1/\alpha} \tag{2}$$

Note that it is possible to keep the Lotka exponent α constant while T is replaced by $T+m$. This is explained in (Egghe & Rousseau, 2006). The argument is based on Egghe (2005, II.2.1).

Table 5
h-indices of Price awardees and recalculated *h*-indices (denoted as h'), based on the assumption that for each of them five highly cited articles are missing

Price awardees	*h*-index	h'	Difference
Braun T.	25	27	2
Egghe L.	13	15	2
Garfield E.	27	29	2
Glänzel W.	18	21	3
Ingwersen P.	13	16	3
Leydesdorff L.	13	15	2
Martin B.	16	19	3
Moed H.F.	18	20	2
Narin F.	27	28	1
Rousseau R.	13	13	0
Schubert A.	18	21	3
Small H.	18	22	4
Van Raan A.F.J.	19	20	1
White H.D.	12	15	3

We want to prove that $h_2 - h_1 \ll m$ or $\Delta h/\Delta T = (h_2 - h_1)/m \ll 1$. In a continuous framework this means that we have to show that $\mathrm{d}h/\mathrm{d}T \ll 1$.

Proposition. *Using the notation explained above we have*: $\mathrm{d}h/\mathrm{d}T \ll 1$.

Proof. $\mathrm{d}h/\mathrm{d}T = \mathrm{d}(T^{(1/\alpha)})/\mathrm{d}T = 1/\alpha T^{((1-\alpha)/\alpha)}$. As $\alpha > 1$, the exponent $(1-\alpha)/\alpha$ is always negative. As T is assumed to be relatively large, this means that $T^{((1-\alpha)/\alpha)} \ll 1$ ($<\alpha$). This shows that $\mathrm{d}h/\mathrm{d}T \ll 1$.

In (Egghe & Rousseau, 2006) we have already shown that the h-index is a concavely increasing function of T, T being the total number of sources (keeping the Lotka exponent α constant). Moreover, $\mathrm{d}h/\mathrm{d}T$ is convexly decreasing with $\lim_{T\to\infty}(\mathrm{d}h/\mathrm{d}T) = \lim_{T\to\infty}(1/\alpha)T^{(1-\alpha/\alpha)} = 0$. These results imply that the influence of missing articles is largest for small T, and tends to zero the larger the number of sources.

Note that in the discrete case we focused on the number of highly cited missing articles, as the other ones clearly have no influence on the value of the h-index. The argument used in the continuous case only uses missing articles, whether or not they are highly cited. Again, this is just a matter of convenience as missing articles that are not highly cited have no influence on the value of the h-index.

In our earlier article (Egghe & Rousseau, 2006) we also derived a formula for the h-index as a function of the total number of items (citations in this article). This relation is, for $\alpha > 2$, given by:

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

where A denotes the total number of items. Clearly, also the number of missed citations does not have a large influence on the value of the h-index. Indeed, $(\mathrm{d}h/\mathrm{d}A) = (1/\alpha)((\alpha-2)/(\alpha-1))^{1/\alpha}A^{((1-\alpha)/\alpha)} \ll 1$, showing that, under the assumption of a fixed α-value, the h-index is stable under the influence of missing citations. □

Remarks.

1. We do not claim that citations always follow a power law, or that adding new articles automatically leads to a new power law with the same α-value. We just say that within this model the analytical results confirm the combinatorial result. Moreover, in this model the influence of missing articles with respect to the number of sources (T) is exactly as one would expect: largest (but still smaller than the number of missed articles) for small T and non-existing for large T.
2. Similar conclusions can be made for the g-index, as introduced by Egghe (2006a, 2006b). Indeed, Egghe has shown that for the power law model $g = ((\alpha-1)/(\alpha-2))^{((\alpha-1)/\alpha)}T^{1/\alpha}$ (Egghe, 2006c). In this model, the difference between the h-index and the g-index is only a factor depending on α. Hence, the g-index, considered as a function of T behaves in the same way as the h-index.

6. Conclusion

Contrary to what one might intuitively expect, a relative small number of missing highly cited publications has only a small influence on the value of the h-index. This is usually the case as shown by the examples of citations following a Zipf distribution, and the h-indices of Price medallists. An analytical model reinforces our argument for missing publications, as well as for missing citations. We conclude that the h-index is resilient to missing articles and to missing citations.

Acknowledgements

Research for this note was performed while the author was a guest of WISE-Lab, Dalian University of Technology and of the National Library of Sciences of CAS (Beijing). He thanks Profs. Liu Zeyuan and Jin Bihui for their hospitality. The author further thanks Prof. Leo Egghe (Hasselt University) for a number of helpful suggestions, improving the obtained results. Research for this article was supported by NSFC Grant Nr. 70373055.

References

Bar-Ilan, J. (2006). *H*-index for Price medallists revisited. *ISSI Newsletter*, *2*(1), 3–5.
Braun, T., Glänzel, W., & Schubert, A. (2005). A Hirsch-type index for journals. *The Scientist*, *19*(22), 8.
Egghe, L. (2005). *Power Laws in the Information Production Process: Lotkaian Informetrics*. Oxford, UK: Elsevier.
Egghe, L. (2006a). How to improve the *h*-index. *The Scientist*, *20*(3), 14.
Egghe, L. (2006b). An improvement of the *H*-index: the *G*-index. *ISSI Newsletter*, *2*(1), 8–9.
Egghe, L. (2006c). Theory and practice of the g-index. *Scientometrics*, *69*, 131–152.
Egghe, L. Dynamic *h*-index: the Hirsch index in function of time. *Journal of the American Society for Information Science and Technology*, in press.
Egghe, L., & Rousseau, R. (2006). An informetric model for the Hirsch index. *Scientometrics*, *69*, 121–129.
Glänzel, W. (2006). On the *h*-index—a mathematical approach to a new measure of publication activity and citation impact. *Scientometrics*, *67*, 315–321.
Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the USA*, *102*(46), 16569–16572.
Liang, L. (2006). *H*-index sequence and *h*-index matrix: constructions and applications. *Scientometrics*, *69*(1), 153–159.
Rousseau, R. (2005). Conglomerates as a general framework for informetric research. *Information Processing and Management*, *41*, 1360–1368.
Rousseau, R. (2006). A case study: evolution of JASIS' *h*-index. E-LIS: ID-code 5430.

Available online at www.sciencedirect.com
ScienceDirect
Journal of Informetrics 1 (2007) 59–67
Journal of INFORMETRICS An International Journal
www.elsevier.com/locate/joi

The source-item coverage of the exponential function

Thierry Lafouge

Laboratoire Ursidoc Université Claude Bernard Lyon 1, 43 Boulevard du 11 novembre 1918, 69622 Villeurbanne Cedex, France

Received 17 July 2006; received in revised form 18 September 2006; accepted 19 September 2006

I dedicate this work to the memory of B. Delattre, brilliant mathematician.

Abstract

Statistical distributions in the production of information are most often studied in the framework of Lotkaian informetrics. In this paper, we recall some results of basic theory of Lotkaian informetrics, then we transpose methods (Theorem 1) applied to Lotkaian distributions by Leo Egghe (Theorem 2) to the exponential distributions (Theorem 3, Theorem 4). We give examples and compare the results (Theorem 5). Finally, we propose to widen the problem using the concept of exponential informetric process (Theorem 6).

Keywords: Exponential function; Mathematical-fitting; Lotkaian informetrics

1. Introduction

Many phenomena studied in informetrics, concerning the production or use of information, can be represented by a triple (source, production function, items) called information production process (IPP) (Egghe, 1990). This consists of a set of sources S, a set of items I, T (respectively, A) denotes the total number of sources, the total number of items, respectively and finally a function of production or use that quantifies the production of the items by the sources. There are several methods for representing these phenomena. In this article, the theory is developed with a size-frequency function f, the most usual form for quantifying this production.

f: $[1, I_{\max}] \to \Re^+$, $f(j)$ describes the density of sources with item density j (in the discrete setting $f(j)$ indicates the number of sources that have produced j items). We assume f is continuous.

$I_{\max}$ indicates the maximal item per source density.

The following two equalities allow us to calculate T and A.

$$T = \int_1^{I_{\max}} f(j)\,\mathrm{d}j \tag{1.1}$$

$$A = \int_1^{I_{\max}} jf(j)\,\mathrm{d}j \tag{1.2}$$

Of course it may happen that T and A are infinite if $I_{\max}$ is infinite. If T is finite we have the following inequality: $0 < T < A$. We denote $\mu = A/T$, the average number of items by source. We have, $\mu > 1$.

E-mail address: Lafouge@univ-lyon1.fr.

1751-1577/$ – see front matter
doi:10.1016/j.joi.2006.09.004

In practice, the production function of an IPP has similar characteristics in very diverse situations of production or use of information:

- authors (sources) write articles (items),
- words in a text (sources) produce occurrences of words in the text (items),
- web pages (sources) contain links (items),
- web sites (sources) are visited (items),
- requests (sources) through a search engine are sent by users (items).

In all quoted examples, if we quantify the production of the items by the sources with a size-frequency function, this one is decreasing with a long tail and a gap between a high number of sources producing few items and a small number of sources producing a lot. In practice, when one determines a best-fitting curve, we must truncate the distribution because, for high frequencies the number of items produced is very low. This characteristic results in the standard deviation often being extremely high compared with the average and is a poor indicator. The statistical distribution mostly used in informetrics is the inverse power function, also called Lotkaian informetric distribution. This distribution is unimodal; it models the information production processes in many of the quoted examples. At present, with Internet, there are many examples for which the data resulting from the Web has been adjusted by such models (Bilke & Peterson, 2001). The zero-truncated generalized inverse Gaussian–Poisson (GIGP) model known and tested over a long time (Burrel & Fenton, 1993) is also used to day to adjust some of this data (Ajiferuke & Wolfram, 2004).

2. Lotkaian informetric distribution

With the preceding notations, a Lotkaian informetric distribution is given:
f: $[1, I_{\max}] \to \Re^+$,where

$$f(j) = \frac{C}{j^{\alpha}}, \quad C > 0 \quad \text{and} \quad \alpha > 1 \tag{2.1}$$

We will limit ourselves to the case where $\alpha > 1$, meaning where T (number of sources) is finite and where the corresponding probability density function is: $f(j) = (\alpha - 1) \times j^{-\alpha}$ $\alpha > 1$, if $I_{\max} = \infty$. Moreover, we know that A is finite if $\alpha > 2$. More generally, f has moments of order n if $\alpha > n$.

The mathematical properties of these functions (Haitun, 1982) have been studied to a great extent. They have often been opposed to the functions modeling Gaussian processes. They have been the subject of a recent work of informetrics (Egghe, 2005), which contains many results. This work has the merit among others of unifying all the work done concerning empirical applications, Lotka, Bradford, Zipf, Mandelbrot, with the mathematical theory of IPP, choosing, as central distribution, the Lotkaian distributions. The coefficient α characterizes the gap between strongly productive sources and those that produce little. Many works (Bookstein, 1990a, b) have shown the strength of the law of Lotka (Lotka, 1926). In addition, the value $\alpha = 2$ plays a key role since we know, according to whether α is smaller or greater than 2, that the representation of Leimkuhler (Rousseau, 1988) has or does not have a turning point; in informetrics the term "Groos droop" (Groos, 1967) is often used.

Finally, these distributions are scale free. A function f is called scale-free if, for every positive constant C, there is a positive constant D such that $f(Cx) = Df(x)$ for all x in the domain of f (Egghe, 2005, p. 27). This property is important when frequencies are observed. It allows us to change scale without changing model. In the main, it justifies the choice of Leo Egghe in his work that we have just quoted. If we impose the scale free property, it implies that some decreasing functions, such as the decreasing exponential function one, are not allowed.

However, the phenomena of obsolescence and growth of quotations on a subject of search in scientific literature (Egghe, 1993) are modeled by exponential processes.

In addition, the often ignored result of Naranan (Naranan, 1971), shows that distributions of Lotkaian type can be deduced under certain conditions from an exponential growth of sources and the number of items produced by these sources. It gives the exponential functions an importance that we cannot neglect. In the previously quoted examples, the temporal parameter plays the role of variable.

Finally, the law of geometrical probability, which is the discrete version of the exponential law, is often used as a rough approximation for modeling the processes of commands or library circulation data (Bagust, 1983).

Thus, all these reasons lead us to adopt a procedure similar to that of Leo Egghe (Egghe, 2005) and to find a mathematical result for the exponential distributions, which is a necessary condition of the same type as Lotkaian distributions.

3. Reminder of some results of basic theory of Lotkaian informetrics

There is a lot of statistical work that consists of verifying the statistical regularities in the variety of examples mentioned in the introduction, and making fittings. Lotkaian distributions play an important role here. However, to my knowledge, few bibliometric researchers use the mathematical theorem, certainly recent, which we remind is a necessary condition for the production of sources covering the items produced. This theorem plays a key role in this article.

Theorem 1. (Egghe, 2005, p. 111)
The following assertions are equivalent, given $A > T > 0$

(i) There exists a function f: $[1,+\infty[\rightarrow \Re^+$ and a finite number $I_{\max} > 1$ such that:

$$T = \int_1^{I_{\max}} f(j)\,\mathrm{d}j$$

$$A = \int_1^{I_{\max}} jf(j)\,\mathrm{d}j$$

(ii) There exists a function f^*: $[1,+\infty[\rightarrow \Re^+$ such that

$$\mu < \frac{\int_1^{\infty} jf^*(j)\,\mathrm{d}j}{\int_1^{\infty} f^*(j)\,\mathrm{d}j}$$

Moreover if (i) or (ii) holds we have, necessarily that $f^* = D \times f$ with $D > 0$, a constant.

We refer readers interested in the demonstration to the reference quoted. What is interesting for us here in the theorem is the implication (ii) $\Rightarrow$ (i).

Leo Egghe applies this theorem to Lotkaian informetric distributions.

Theorem 2. (Egghe, 2004)
Let $0 < T < A < \infty$ be given. Let $\alpha > 1$ and a number $I_{\max} > 1$.
If $I_{\max}$ is infinite

(i) If the inverse power function as in (2.1) satisfies (1.1) and (1.2) if we have

$$\alpha = \frac{2\mu - 1}{\mu - 1} \tag{3.1}$$

and

$$C = \frac{A}{\mu - 1} \tag{3.2}$$

which implies $\alpha > 2$ if $A < \infty$.

If $I_{\max}$ is finite (the general case)

(ii) If $\alpha \leq 2$ then there always exists a number $I_{\max} > 1$ such that (2.1) satisfies (1.1) and (1.2).
(iii) If $\alpha > 2$ the conclusion (i) is valid if and only if

$$\mu < \frac{\alpha - 1}{\alpha - 2} \tag{3.3}$$

We will follow exactly the same procedure for the exponential functions (Theorem 3 and Theorem 4), then compare the results (Theorem 5).

4. Exponential distribution

4.1. Theoretical results

With the preceding notations, an exponential function: g: $[1, I_{\max}] \to \Re^+$ is given where

$$g(j) = C\,\mathrm{e}^{-\alpha(j-1)} \quad \text{with} \quad C > 0 \quad \text{and} \quad \alpha > 0 \tag{4.1}$$

We can also write it in an equivalent form:

$$g(j) = Ca^{-j} \quad \text{with} \quad C > 0 \quad \text{and} \quad a > 1$$

We will use the first form here. The corresponding probability density function is:

$$g(j) = \alpha\,\mathrm{e}^{-\alpha(j-1)} \quad \text{with} \quad \alpha > 0 \quad \text{and} \quad I_{\max} = \infty$$

As for the power function, g has as a maximum value for 1.
Unlike the inverse power function, a decreasing exponential function has moments of order n whatever the n positive.

Lemma. If we call $A(n) = \int_1^\infty j^n\,\mathrm{e}^{-\alpha(j-1)}\,\mathrm{d}j$ the moment of order n divided by α of an exponential function, we have

$$A(n) = \sum_{p=0}^{p=n-1} \frac{n!}{(n-p)!}\frac{1}{\alpha^{p+1}} + \left(n!\frac{1}{\alpha^{n+1}}\right), \quad n \geq 0;$$

Proof. An integration by part gives:

$$A(n) = \frac{1}{\alpha} + \frac{n}{\alpha}A(n-1)$$

with $A(0) = 1/\alpha$. We show by recurrence:

$$A(n+1) = \frac{1}{\alpha} + \frac{(n+1)}{\alpha}A(n) = \frac{1}{\alpha} + \frac{(n+1)}{\alpha}\left(\sum_{p=0}^{p=n-1}\frac{n!}{(n-p)!}\frac{1}{\alpha^{p+1}} + n!\frac{1}{\alpha^{n+1}}\right)$$

$$= \frac{1}{\alpha} + \left(\sum_{p=0}^{p=n-1}\frac{(n+1)!}{(n-p)!}\frac{1}{\alpha^{p+2}} + (n+1)!\frac{1}{\alpha^{n+1}}\right) = \sum_{p=0}^{p=n}\frac{(n+1)!}{(n+1-p)!}\frac{1}{\alpha^{p+1}} + (n+1)!\frac{1}{\alpha^{n+2}}$$

More particularly we obtain:

$$A(1) = \frac{1}{\alpha} + \frac{1}{\alpha^2} \quad \square \tag{4.2}$$

Problem. What are the conditions, given A and T $(0 < T < A)$, for the existence of an exponential function g as in (4.1) which verifies:

$$\int_1^{I_{\max}} g(j)\,\mathrm{d}j = T \quad \text{and} \quad \int_1^{I_{\max}} jg(j)\,\mathrm{d}j = A?$$

We separate the study into two cases.

(1) $I_{\max}$ is infinite

Theorem 3. Let $0<T<A$ be given. The exponential function defined in (4.1) satisfies the following conditions:

$$\int_1^{\infty} g(j)\,\mathrm{d}j = T; \qquad \int_1^{\infty} jg(j)\,\mathrm{d}j = A$$

if

$$\alpha = \frac{T}{A-T} = \frac{1}{\mu-1} \tag{4.3}$$

$$C = \frac{T^2}{A-T} \tag{4.4}$$

Proof. By solving the integrals above (see (4.2)), we obtain:

$$\int_1^{\infty} C\,\mathrm{e}^{-\alpha(j-1)}\,\mathrm{d}j = \frac{C}{\alpha}$$

and

$$\int_1^{\infty} Cj\,\mathrm{e}^{-\alpha(j-1)}\,\mathrm{d}j = C\left(\frac{1}{\alpha^2}+\frac{1}{\alpha}\right)$$

We then deduce the desired results ((4.3), (4.4)) by solving the two equations:

$$T = \frac{C}{\alpha}$$

and

$$A = \frac{C\alpha + C}{\alpha^2}.$$

As for a Lotkaian distribution, α only depends on μ. When fitting, the formulas (4.3), (4.4) give a rough estimate of the parameters of the exponential function (4.1). □

(2) $I_{\max}$ is finite

Theorem 4. Let $A>T>0$ be given. Thus, $\alpha>0$, there is still $I_{\max}>1$ finite and an exponential function defined by (4.1) verifying the two conditions:

$$\int_1^{I_{\max}} g(j)\,\mathrm{d}j = T \text{ and } \int_1^{I_{\max}} jg(j)\,\mathrm{d}j = A$$

if the inequality

$$\mu < 1 + \frac{1}{\alpha} \tag{4.5}$$

holds.

Proof. According to the preceding lemma, we have:

$$\frac{\int_1^\infty jC\,\mathrm{e}^{-\alpha(j-1)}\,\mathrm{d}j}{\int_1^\infty C\,\mathrm{e}^{-\alpha(j-1)}\,\mathrm{d}j} = \frac{(1/\alpha^2)+(1/\alpha)}{(1/\alpha)} = 1+(1/\alpha)$$

The assertion (ii) of Theorem 1 allows us to conclude. It will be noticed that the result is valid for any value $\alpha>0$. In particular, the value $\alpha=2$, unlike the Lotkaian distribution, is not a key value.

Construction of *g*

Now its existence is proven, we must show how to build it. To simplify the notations we put $x=I_{\max}$.

$$\int_1^x C\,\mathrm{e}^{-\alpha(j-1)}\,\mathrm{d}j = T \Rightarrow \frac{T}{C} = \frac{1-\mathrm{e}^{-\alpha(x-1)}}{\alpha}$$

$$\int_1^x jC\,\mathrm{e}^{-\alpha(j-1)} = A \Rightarrow \frac{A}{C} = \frac{1-x\,\mathrm{e}^{-\alpha(x-1)}}{\alpha} + \frac{1-\mathrm{e}^{-\alpha(x-1)}}{\alpha^2}$$

We suppose $x\neq 1$. By eliminating C we deduce the following equation:

$$\frac{A\alpha}{T} = \frac{\mathrm{e}^{-\alpha(x-1)}(-1-x\alpha)+\alpha+1}{1-\mathrm{e}^{-\alpha(x-1)}}$$

thus,

$$\mu\alpha - \frac{\mathrm{e}^{-\alpha(x-1)}(-1-x\alpha)+\alpha+1}{1-\mathrm{e}^{-\alpha(x-1)}} = 0 \tag{4.6}$$

Unlike the case where x is infinite there are many values $\alpha>0$ where the preceding equation has solutions. We consider α as a parameter of the Eq. (4.6), $\alpha>0$. We solve this equation in x, by the iterative method, using the MAPPLE 4.0 software for example. Then we calculate C,

$$C = \frac{\alpha T}{1-\mathrm{e}^{-\alpha(x-1)}} \quad \square \tag{4.7}$$

We have just seen a necessary condition for an exponential function to produce a given number of items with a given number of sources. We shall see that if this necessary condition holds, then it also holds for a Lotkaian distribution. More precisely, we have the following result:

Theorem 5. Let $A>T>0$ be given. Let $\alpha>1$. If there is a number $I_{\max}>1$ such that (4.1) satisfies (1.1) and (1.2), then it is also valid for (2.1).

Proof. If $I_{\max}=\infty$ it is still true according to the results (i) of Theorem 2.

If $I_{\max}$ is finite we have two cases.

(i) $\alpha\leq 2$, we know according to the result (ii) of Theorem 2 that it is also true.
(ii) $\alpha>2$, we know according to Theorem 4 that (4.5) is true, thus, $\alpha<[1/(\mu-1)]$, then $1/(\mu-1)<[2(\mu-1)/\mu-1]$ thus, the inequality $\alpha<[(2\mu-1)/(\mu-1)]$ is true, thus, the inequality (3.3) is true, the assertion (iii) of Theorem 2 then allows us to conclude. $\square$

4.2. Examples

(1) A=10,000, T=5000 thus, $\mu=2$ and $\alpha=0.5$. The inequality (4.5) is demonstrated. We must then solve the Eq. (4.6):

$$1-\frac{\mathrm{e}^{-0.5(x-1)}(-1-0.5x)+1.5}{1-\mathrm{e}^{-0.5(x-1)}} = 0.$$

We obtain the solution $x=3.512$, then according to (4.7) we have $C\approx 8778$.

The desired exponential function is: $g(j) = 8778\,\mathrm{e}^{-0.5(j-1)}$.

(2) A=10,000, T=7000 thus, $\mu = 1.43$ and $\alpha = 2$. The inequality (4.5) is demonstrated. We must then solve the Eq. (4.6):

$$2.86 - \frac{\mathrm{e}^{-2(x-1)}(-1-2x)+3}{1-\mathrm{e}^{-2(x-1)}} = 0.$$

We obtain the solution $x = 2.58$ then according to (4.7) we have $C \approx 1462$

The desired exponential function is: $g(j) = 1462\,\mathrm{e}^{-2(j-1)}$.

Note

The results in Section 4 could be considered as a "mathematical fitting" method for exponential function, as opposed to statistical fitting.

5. Perspectives: exponential informetric process

In the article (Lafouge & Prime Claverie, 2005) we define an exponential informetric process in terms of an exponential function and an effort function where the average quantity supplied by the sources, to produce all the items is finite. More precisely, a set of functions, denoted EF.

EF = $\{h: [1, \infty[\to \Re^+$: increasing, continuous, and not majorized$\}$

$h \in$ EF an effort function is then any element of EF.

We call exponential informetric process the size-frequency function $\nu(h)$:

$$\nu(h)(j) = C\,\mathrm{e}^{-h(j)} \quad C > 0 \tag{5.1}$$

where the following quantity

$$F = \int_1^\infty \nu(h)(j)h(j)\,\mathrm{d}j \tag{5.2}$$

is finite, F corresponds to the quantity of effort produced by $\nu(h)$.

Note

The total number of sources T, $T = \int_1^\infty \nu(h)(j)\,\mathrm{d}j$ is finite and we have the inequality $\infty > F > T > 0$.

Examples

The respective functions of effort, $h(j) = \alpha \times \ln(j)$, $\alpha > 1$ and $h(j) = \alpha(j-1)$, $\alpha > 0$ correspond to the inverse power function $f(j) = C/j^\alpha$ and to the exponential function $g(j) = C \times \mathrm{e}^{-\alpha(j-1)}$, studied previously.

Problem. What are the conditions, given the quantity of effort F, the number of sources T, for the existence of an exponential informetric process $\nu(h)$ as in (5.1), where $I_{\max}$ is a number > 1, which verifies $F = \int_1^{i_{\max}} \nu(h)(j)h(j)\,\mathrm{d}j$ and $T = \int_1^{i_{\max}} \nu(h)(j)\,\mathrm{d}j$?

We will limit ourselves to the case where the respective functions of effort correspond to the inverse power function (2.1) and to the exponential function (4.1), and where $I_{\max} = \infty$.

Theorem 6. Let $F > T > 0$ be given:

(i) the exponential informetric process as in (5.1) where $h(j) = \alpha(j-1)$,$\alpha > 0$ satisfies the following conditions:

$$T = \int_1^{\infty} C\,\mathrm{e}^{-\alpha(j-1)}\,\mathrm{d}j$$

$$F = \int_1^{\infty} C\,\mathrm{e}^{-\alpha(j-1)}\alpha(j-1)\,\mathrm{d}j$$

if

$$T = F = C/\alpha \tag{5.3}$$

(ii) the exponential informetric process as in (5.1) where $h(j) = \alpha \times \ln(j)$, $\alpha > 1$ satisfies the following conditions:

$$T = \int_1^{\infty} C\frac{1}{j^{\alpha}}\,\mathrm{d}j$$

$$F = \int_1^{\infty} C\frac{1}{j^{\alpha}}\alpha\ln(j)\,\mathrm{d}j$$

if

$$\alpha = \frac{F}{F-T} \tag{5.4}$$

$$C = \frac{T^2}{F-T} \tag{5.5}$$

Proof.

(1) By (4.1) $T = C/\alpha$. An integration by part give: $F = C/\alpha$
(2) By (2.1) $T = \frac{C}{\alpha-1}$

$\int_1^{\infty} \frac{1}{j^{\alpha}} \ln(j)\,\mathrm{d}j = -\frac{1}{\alpha-1}\int_1^{\infty} \ln(j)\,\mathrm{d}\left(\frac{1}{j^{\alpha-1}}\right)$, an integration by part give: $\int_1^{\infty} \frac{1}{j^{\alpha}} \ln(j)\,\mathrm{d}j = \frac{1}{(\alpha-1)^2}x$ thus, $F = \frac{C\alpha}{(\alpha-1)^2}$. We then deduce the desired results (5.4) and (5.5) by solving the two equations:

$$T = \frac{C}{\alpha-1}$$

and

$$F = \frac{C\alpha}{(\alpha-1)^2} \quad \square$$

The case where $I_{\max}$ is finite is an open problem.

References

Ajiferuke, I., & Wolfram, D. (2004). Informetric modelling of Internet search and browsing characteristics. *The Canadian Journal of Information and Library Science*, *28*(1), 1–16.

Bagust, A. (1983). A circulation model for busy public libraries. *Journal of Documentation*, *39*(1), 24–37.

Bilke, S., & Peterson, C. (2001). Topological properties and metabolic networks. *Physical Reviews E*, *6403*(3), 76–80.

Bookstein, A. (1990a). Informetric distribution, Part 1: unified overview. *Journal of the American Society for Information Science*, *41*(5), 368–375.

Bookstein, A. (1990b). Informetric distribution, Part 2: resilience to ambiguity. *Journal of the American Society for Information Science*, *41*(5), 376–385.

Burrel, Q. L., & Fenton, M. R. (1993). Yes, the GIGP really does work and is workable. *Journal of the American Society for Information Science*, *44*(2), 61–69.

Egghe, L. (1990). On the duality of informetric systems with applications to the empirical law. *Journal of Information Science*, *16*, 17–27.

Egghe, L. (1993). On the influence of growth on obsolescence. *Scientometrics*, *27*(1), 195–214.

Egghe, L. (2004). The source-item coverage of the Lotka function. *Scientometrics*, *61*(1), 103–115.
Egghe, L. (2005). *Power laws in the information production process: Lotkaian informetrics*. Elsevier.
Groos, A. V. (1967). Bradford's law and the Keenan–Atherton data. *American Documentation*, *18*, 46.
Haitun, S. D. (1982). Stationary scientometric distributions. *Scientometrics* no. 4, Part I, 5–25, Part II, 89–104, Part III, 181–194.
Lafouge, T., & Prime Claverie, C. (2005). Production and use of information. Characterization of informetric distributions using effort function and density function exponential informetric process. *Information Processing and Management*, *41*, 1387–1394.
Lotka, A. J. (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Science*, 292–306.
Naranan, S. (1971). Bradford Law of bibliography of science an interpretation. *Nature*, *227*(5258), 631–632.
Rousseau, R. (1988). Lotka's law and its Leimkuhler representation. *Library Science with a Slant to Documentation and Information Studies*, *25*(3), 150–178.

Available online at www.sciencedirect.com
ScienceDirect
Journal of Informetrics 1 (2007) 123–130
Journal of INFORMETRICS
An International Journal
www.elsevier.com/locate/joi

A rational indicator of scientific creativity

José M. Soler

Departamento de Física de la Materia Condensada, C-III, Universidad Autónoma de Madrid, E-28049 Madrid, Spain

Received 31 August 2006; received in revised form 31 October 2006; accepted 31 October 2006

Abstract

A model is proposed for the creation and transmission of scientific knowledge, based on the network of citations among research articles. The model allows to assign to each article a non-negative value for its creativity, *i.e.* its creation of new knowledge. If the entire publication network is truncated to the first neighbors of an article (the n references that it makes and the m citations that it receives), its creativity value becomes a simple function of n and m. After splitting the creativity of each article among its authors, the cumulative creativity of an author is then proposed as an indicator of her or his merit of research. In contrast with other merit indicators, this creativity index yields similar values for the top scientists in two very different areas (life sciences and physics), thus offering good promise for interdisciplinary analyses.

Keywords: Citation analysis; Citation network; Knowledge flow; Research merit; Scientific creativity

1. Introduction

Evaluating the scientific merit and potential, of tenure and professorship candidates, is perhaps the most critical single activity in the academic profession. In countries and institutions with a long scientific tradition, selection committees are generally well trained and trusted to balance wisely the vast variety of factors that may influence the decision, in the sense of optimizing the long-term scientific output. In less established environments, decisions are frequently perceived as arbitrary, and the use of objective indicators and procedures may be necessary to obtain a wide consensus (Moed, 2005).

The most traditional indicator of research output, the number of published papers, has been progressively substituted by the number of citations received by those papers, when this impact indicator has become widely available and easy to obtain (Garfield, 1964; ISI-Thomson, 2006). Different combinations of both magnitudes have been proposed (Rinia, van Leeuwen, van Vuren, & van Raan, 1998) like those in the SPIRES database (SPIRES, 2006). The field has been recently revitalized by the proposal by Hirsch (2005) of yet another combination, the so-called h index, which has gained a rapid popularity, partly because the ISI-Thomson Web of Knowledge database (ISI-Thomson, 2006) provides a handy tool to sort articles by their number of citations. Apart from that comparative handiness, there is little objective evidence for the relative advantages of different indexes, which are generally motivated in terms of "impact" or "influence". However, it must not be forgotten that the task of a scientist is to create useful knowledge (in its broadest sense), not merely to produce an impact. It is therefore desirable to derive some rational measure of the magnitude and quality

E-mail address: jose.soler@uam.es.

1751-1577/$ – see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.joi.2006.10.004

of research output, rooted in a plausible model of the creation and transmission of scientific knowledge (Rinia, van Leeuwen, Bruins, van Vuren, & van Raan, 2002).

2. Creativity model

Basic scientific knowledge, as opposed to technological or industrial knowledge, is created by the minds of scientists and expressed almost exclusively as research articles. The knowledge is transmitted to other scientists, who read previous articles and acknowledge this transmission in the form of references (in what follows, I will call *references* of an article those made to previous papers, and *citations* those received from posterior papers). Thus, the output knowledge of an article comes partly from previous work, which is simply transmitted, and partly from the creation of new knowledge by the authors. However, there are many possible reasons why references are made (Cozzens, 1989; Garfield, 1964; Gilbert, 1977; Merton, 1968). Furthermore, some of the references of an article may be more important than others. Thus, it is rather uncertain to what extent a given reference reflects the use of previous knowledge. Therefore, in the present model, I will simply assume that each reference reflects the transmission of a different non-negative value x_{ij} of knowledge, with probability $P(x_{ij})$, from the cited article i to the citing article j. The maximum entropy principle (Tribus, 1969) dictates that, in the absence of any *a priori* information, other than the average value $\langle x\rangle = 1/\alpha$, the probability is given by $P(x) = \alpha\, \mathrm{e}^{-\alpha x}$.

Consider the network formed by all published papers connected by their citations. The growth, connectivity, and statistical properties of this and similar networks have been the subject of much recent work (Albert & Barabasi, 2002; Redner, 2005). To model the flow of knowledge on this supporting network (Rinia et al., 2002), we may assign random flow numbers x_{ij} to all citations, with probability $P(x_{ij})$. Flow conservation implies that the articles' knowledge-creation values c_i (that I will simply call creativities) obey

$$c_i = \sum_j x_{ij} - \sum_k x_{ki} \tag{1}$$

I will discard negative knowledge as meaningless.[1] Thus, I will require that $c_i \geq 0\ \forall i$, and reject the sets $\{x_{ij}\}$ that violate this condition.[2,3,4] The final values c_i will then be averages over all valid sets $\{x_{ij}\}$, with a relative weight $P(\{x_{ij}\}) \propto \exp(-\alpha \sum_{ij} x_{ij})$.

Some attention must be paid to the definition of knowledge that is being used. It might seem that all the knowledge created by an article must be present already when it is published. However, this would make it difficult to judge the relative importance of the knowledge created by different papers. Therefore, I rather consider the amount of "used knowledge" (and therefore useful). The situation is very similar in commercial software development: the economic value of a computer library does not materialize when it is written, but when licenses of it are sold, presum-

[1] An ironic observer might object to this assumption, arguing that many articles contribute only to confusion, and that some citations are in fact critical. I find this questionable, since most readers will filter efficiently this "negative" knowledge, simply ignoring it. Also, even wrong ideas can stimulate new valid ones. In any case, critical references cannot be easily distinguished from positive ones, but their average effect might be taken into account by renormalizing the mean flow value $\langle x\rangle$.

[2] Since new, non-negative knowledge is created in every article and transmitted to the future, the total flow of knowledge must increase with time. Such an increase may be absorbed in three ways: by an increase in the number of articles published per year, by an increase in the number of references per article, and by an increase in the average flow per citation $\langle x\rangle$. The increase of the rate of publications is indeed a large effect, while that of citations per paper is much weaker, if positive at all. In any case, it is not clear whether those two effects combined can fully account for the transmission of the new knowledge predicted by the model. Thus, it may be necessary to adjust self-consistently a function $\alpha(t)$, of time t. In this work, I have taken $\alpha = \text{constant} = 1$.

[3] Some of the basic scientific knowledge "leaks" out of the academic research literature in various forms: as knowledge absorbed by scientists who read the articles but do not cite them; as established knowledge transmitted to textbooks and no longer cited in research articles (oblivion by incorporation); as technological knowledge translated to patents, that may cite the literature but that are not included in databases of basic research (Narin, Hamilton, & Olivastro, 1997); and as industrial knowledge translated to unpublished manufacture methods and products. It seems reasonable to assume that this "hidden" flow of knowledge is proportional on average to the "visible" flow shown by citations. Therefore, in order to account for the hidden flow, we may multiply the visible output flow of each article (first term of Eq. (1)) by a factor $(1+\gamma)$, where γ is a phenomenological adjustable parameter. In the simplified model of this work I have taken $\gamma = 0$.

[4] The boundary problem posed by recent papers, that have had no time to transmit their knowledge, may be addressed by not imposing flow conservation on them, or by assigning to them an average number of additional expected citations, that will be a decreasing function of their age. In any case, it is clear that any figure of merit based on citations will not be as reliable for very recent papers as for old ones.

ably to create new software. Similarly, I am counting every "copy" of the knowledge, used in every new paper that cites it.

This economic analogy can be pursued further. Thus, I will argue that, to a large extent, the citation network acts as a market in which citing and cited articles act, respectively, as buyers and sellers of scientific knowledge. At first sight, it might seem that the price of existing knowledge is zero, since it costs nothing to read and cite a published article. However, there are practical limits to the number of references that can be made. These limits are dictated by journal conditions, by the community citation practices, and, most importantly, by the authors limited time to read new articles, and memory to remember them. Therefore, authors are forced in practice to select a limited number of references, that are most relevant to their works. As in any market, the limited resources determine the "prices", in this case the flows of knowledge assigned to each citation.

Some of the general qualitative features of the model, as an indicator of research merit, may be expected *a priori*: articles with less citations than references will have a positive but small creativity value; articles with a large output (very cited) and a small input (not many references) will have the largest creativities; in contrast, the merit of review articles will be much more moderate than that shown by their raw impact factor (citation count); the differences between the creativities of authors in very large and active fields (with large publication and citation rates), and those in smaller and less active fields, will be largely attenuated, as compared to other merit indicators, since the basic measure is the difference between citations and references, which should be roughly zero in all fields; self-citations will be largely discounted, since they will count both as a negative contribution (to the citing paper) and a positive one (to the cited paper); citations received from a successful article (*i.e.* a very cited one itself) will be more valuable than those made by a poorly cited one (Chen, Xie, Maslov, & Redner, 2006; Pinski & Narin, 1976). In particular, citations by uncited papers will add no value at all, since no knowledge can flow through them; more generally, articles that generate a divergent citation tree (*e.g.* the DNA paper of Watson and Crick) will have a large creativity, while those leading ultimately to a dead end (*e.g.* the cold fusion paper of Fleischmann and Pons) will have a small one, even if they had the same number of direct citations.

3. Simplified model

The quantitative analysis of the model presented above is an interesting challenge that will be addressed in the future. In this work, I am rather interested in simplifying the model to allow the easy generation of a practical indicator of merit of research. The simplified model will keep many of the general features discussed above, though not all (in particular, it will loose the last two properties mentioned above). Thus, I propose to truncate the citation network beyond the first neighbors of any given paper, *i.e.* to consider only its n references and m citations, and to impose the conservation of flow, Eq. (1), only in the central node i. The average value $\langle x\rangle$ can be used as a convenient unit of knowledge, so that $\alpha = 1$ and $P(x) = \mathrm{e}^{-x}$. The probability that an article, with n references and m citations, has a creativity c is then, for $n, m > 0$:

$$P(c|n,m) = N^{-1} \int \cdots \int_0^\infty \mathrm{d}x_1 \cdots \mathrm{d}x_n \, \mathrm{d}y_1 \cdots \mathrm{d}y_m \, \delta(c + x - y) \, \mathrm{e}^{-x-y} \tag{2}$$

with $x = \sum_{i=1}^{n} x_i$ and $y = \sum_{j=1}^{m} y_j$, where x_i are the input flows (references) and y_j are the outputs (citations). $\delta(x)$ is Dirac's delta function and N is a normalization factor given by

$$N = \int \cdots \int_0^\infty \mathrm{d}x_1 \cdots \mathrm{d}x_n \, \mathrm{d}y_1 \cdots \mathrm{d}y_m \, \theta(y - x) \, \mathrm{e}^{-x-y} \tag{3}$$

where $\theta(x)$ is the step function. Using a convenient change of variables, the integrals can be evaluated as

$$N = \int_0^\infty \int \frac{\mathrm{d}x \, \mathrm{d}y \, x^{n-1} y^{m-1}}{(n-1)!(m-1)!} \theta(y - x) \, \mathrm{e}^{-x-y} \tag{4}$$

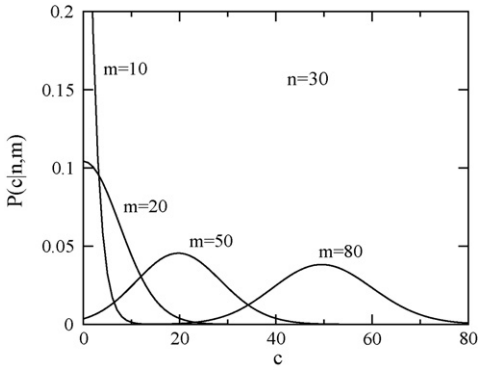


Fig. 1. Probability that an article, that has made $n = 30$ references and has received m citations, has created a value c of scientific knowledge. It was obtained from Eq. (6).

$$P(c|n,m) = N^{-1} \int_0^\infty \int \frac{\mathrm{d}x\,\mathrm{d}y\, x^{n-1} y^{m-1}}{(n-1)!(m-1)!} \delta(c+x-y)\, \mathrm{e}^{-x-y} \tag{5}$$

The result is

$$P(c|n,m) = \frac{n\,\mathrm{e}^{-c}}{n+m-1} \frac{{}_1F_1(1-m, 2-n-m; 2c)}{{}_2F_1(1, 1-m; 1+n; -1)} \tag{6}$$

where ${}_1F_1$ and ${}_2F_1$ are hypergeometric functions, which can be expanded as a finite series (Gradshteyn & Rydhik, 1980). Fig. 1 shows some typical probability distributions.

The average value of c,

$$c(n,m) = \int_0^\infty \mathrm{d}c\, c\, P(c|n,m), \tag{7}$$

is, for $n, m > 0$:

$$c(n,m) = \frac{\sum_{k=0}^{m-1} ((n+m-2-k)!/(m-1-k)!)(k+1)2^k}{\sum_{k=0}^{m-1} ((n-1)!(n+m-1)!)/((n+k)!(m-1-k)!)}. \tag{8}$$

It is represented in Fig. 2 for some typical values of n and m. As expected, $c(n, m)$ increases with m and it decreases with n. It obeys $c(0, m) = m$, $c(n, 0) = 0$, $c(n, 1) = 1$, and $c(n, m) \geq \max(1, m-n)\ \forall m > 0$. For the present purposes, a

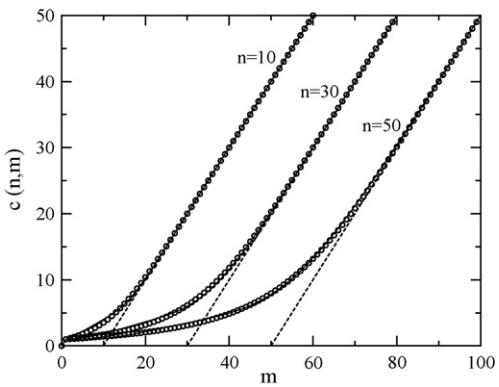


Fig. 2. Circles: mean creation of knowledge (creativity) of an article with n references and m citations, calculated from Eq. (8) (in units of the mean transmission of knowledge reflected by one reference). Solid lines: fits given by Eq. (9). Dashed lines: $m - n$.

reasonably accurate fit is, for $m > 0$:

$$c(n, m) \simeq m - n + \frac{n}{A\,\mathrm{e}^{az} + B\,\mathrm{e}^{bz}} \tag{9}$$

where $z = (m-1)/(n+5)$, $A = 0.986$, $B = 0.014$, $a = 1.08$, and $b = 6.3$. The accumulated creativity of an author with N_p published papers is then defined as

$$C_\mathrm{a} = \sum_{i=1}^{N_\mathrm{p}} \frac{c(n_i, m_i)}{a_i} \tag{10}$$

where a_i is the number of authors of paper i. Notice that, being positive and cumulative, C_a can only increase with time and with the number of published papers.

In order to find in practice the creativity of an author (among many other merit indicators), one can follow these steps: (1) download the programs *filter* and *merit* from this author's web page (Soler, 2006), and compile them if necessary. (2) Perform a "General search" in the ISI-Thomson Web of Science database (ISI-Thomson, 2006) for the author's name, using the appropriate filters. (3) Select the required records. Usually the easiest way is to check "Records from 1 to last_one" and click on "ADD TO MARKED LIST" (if you find too many articles, you may have to mark and save them by parts, say (1–500) → file1, (501–last_one) → file2); (4) click on "MARKED LIST". (5) Check the boxes "Author(s)", "Title", "Source", "keywords", "addresses", "cited reference count", "times cited", "source abbrev.", "page count", and "subject category". Do not check "Abstract" nor "cited references", since this would slow down considerably the next step. (6) Click on "SAVE TO FILE" and save it in your computer. (7) Click on "BACK", then on "DELETE THIS LIST" and "RETURN", and go to step 2 to make another search, if desired. (8) If you suspect that there are two or more authors with the same name, use the *filter* program to help in selecting the papers of the desired author. (9) Run the *merit* program to find the merit indicators. Mind for hidden file extensions, possibly added by your navigator, when giving file names in this and previous step.

4. Results and discussion

Table 1 shows several indexes of merit of top scientists in life sciences and physics, taken from Hirsch's selection (Hirsch, 2005). It may be seen that the h index of all biologists is larger than that of all physicists, and their average number of publications and citations is 1.5–2.5 times larger. In contrast, the two creativity distributions are remarkably similar, with averages that differ only ∼15%, well below the standard deviation of both distributions. This offers the promise of direct interdisciplinary comparisons, without any field normalization, a highly desirable characteristic of any index of merit.

Although it is a natural consequence of the idea of knowledge flow, the fact that the references of an article will result in lowering the merit assigned to it, is admittedly striking. It is thus appropriate to recognize that this is partly due to a deliberate intent of measuring creativity rather than productivity (or, in economic terms, added value rather than sales). To illustrate the point, imagine that two scientists, Alice and Bob, address independently an important and difficult problem in their field. Bob takes an interdisciplinary approach and discovers that a method developed in a different field just fits their need. Simultaneously, Alice faces the problem directly and re-invents the same method by herself (thus making less references in her publication and achieving a higher creativity index).[5] All other factors being equal, both papers will receive roughly the same number of citations, since they transmit the same knowledge to their field. But, while it may be argued that Alice's work was more creative in some sense, it might also be argued that Bob's approach is better, since it additionally shows the relationship with a different field. Thus, although I believe that C_a genuinely correlates with scientific creativity, I would not argue that this is necessarily the most valuable research ability in general.

Eventually, the usefulness of different merit indicators will depend on how well they correlate with real human-made selections (Cole & Cole, 1971; Rinia et al., 1998). Thus, Table 1 shows also a "productivity index" P_a (not a probability), given by the author's share of the citations received by her/his papers. Notice that, in the model proposed,

[5] This is somewhat hypothetical since good citation practice (frequently enforced by the referees) requires that previous relevant work is cited, independently of whether it was actually used.

Table 1
Several merit indicators of the 10 most cited scientists in life sciences and physics (Hirsch, 2005)

Name	N_p	N_c (10^3)	h	P_a (10^3)	C_a (10^3)
B. Vogelstein	447	144.4	154	34.1	32.0
S.H. Snyder	1144	138.3	194	48.2	38.9
S. Moncada	693	106.2	145	32.5	27.8
P. Chambon	987	98.1	153	23.0	17.7
R.C. Gallo	1247	95.9	154	17.9	13.8
D. Baltimore	657	95.3	162	33.0	28.2
R.M. Evans	428	78.8	130	21.2	18.3
T. Kishimoto	1621	77.5	134	14.6	10.2
C.A. Dinarello	992	74.3	138	26.3	19.2
A. Ullrich	615	73.0	122	13.6	10.9
Average	**883**	**98.2**	**149**	**26.4**	**21.7**
Standard deviation	364	24.1	19	10.1	9.1
P.W. Anderson	342	56.7	96	39.1	36.9
A.J. Heeger	999	53.5	109	14.2	10.3
E. Witten	254	53.1	111	39.9	35.9
S. Weinberg	444	38.8	88	32.7	29.3
M.L. Cohen	625	37.4	94	14.3	10.6
M. Cardona	1096	37.0	88	12.8	7.8
A.C. Gossard	918	34.3	92	7.4	5.8
P.G. deGennes	358	32.6	80	26.7	23.9
M.E. Fisher	446	29.8	88	19.0	14.3
G. Parisi	469	24.9	75	12.2	9.9
Average	**595**	**39.8**	**92**	**21.8**	**18.5**
Standard deviation	286	10.4	11	11.3	11.3

N_p, number of papers published; N_c, number of citations received by those papers; h, number of papers with h or more citations (Hirsch index) (Hirsch, 2005); P_a, author's knowledge-productivity index, $P_a = \sum_{i=1}^{N_p} m_i/a_i$, where a_i and m_i are the number of authors and of citations received by paper i; C_a, author's creativity index, Eq. (10). The data were obtained in April 2006.

Table 2
Several indicators of some of the main multidisciplinary, review and non-review physics journals

Journal	N_p	N_r/N_p	N_c/N_p	C/N_p	IF
Nature	3676	10	67	59	28.8
Science	2449	14	74	63	24.4
PNAS	2133	31	112	84	9.8
Rev. Mod. Phys.	20	284	327	160	13.4
Adv. Phys.	8	391	149	18	12.7
Surf. Sci. Rep.	5	159	61	3	10.3
Rep. Prog. Phys.	29	198	90	32	6.2
Phys. Rep.	81	166	90	22	5.6
Phys. Rev. Lett.	1904	18	59	44	6.0
Phys. Rev. D	1049	27	23	11	3.9
Nucl. Phys. B	620	37	42	24	3.3
Appl. Phys. Lett.	1819	13	34	26	3.3
J. Chem. Phys.	2040	37	37	16	3.1
Phys. Rev. B	3488	27	35	18	2.8

N_p, number of "papers" (documents) published in year 1990, in all the sections included in the Science Citation Index database; N_r, number of references made by those papers; N_S, number of citations received by those papers until May 2006 (August 2006 for *PNAS*); C, sum of the creativities, Eq. (8), of those papers, $C = \sum_{i=1}^{N_p} c(n_i, m_i)$; IF, impact factor in 1998 (center of the period 1990–2006), as defined by the Journal of Citation Reports (ISI-Thomson, 2006). For the non-review physics journals (last group), the indicators (other than N_p and IF) have been obtained from a random sample of their N_p papers, rather than from the whole set.

N_c is the total output flow of knowledge from the author's papers, while P_a is her/his share of it. It may be seen that P_a also allows reliable interdisciplinary comparisons. It may be concluded that the main difference between the two communities is the larger average number of authors per article in the life sciences, which is taken into account in both P_a and C_a, but not in the other indexes.

Knowledge-productivity and creativity indicators can be used also for groups, institutions, or journals. Thus, Table 2 shows them for some leading journals. As expected, most review journals have considerably smaller creativities than productivities (dramatically smaller in some cases). Still, *Reviews of Modern Physics* has the largest creativity index of all the journals studied, showing that collecting, processing, and presenting knowledge in a coherent way can by itself create much *new* useful knowledge.

Finally, in a world of strong competition for positions and founds, a negative merit assignment to references might result in a tendency to reduce them below what would be scientifically desirable and professionally fair. A possible solution is to use, in Eq. (8), a fixed value of n (equal to the journal reference intensity, *i.e.* the average number of references per article in that journal), to calculate the creativities for competitive-evaluation purposes. This would spoil a few desirable properties of the model (like the discount of self-citations), but most of its effects would probably be rather mild, since the number of references per paper has a much smaller variance than the number of citations. Thus, the root mean squared difference between the creativities of Table 1, calculated using the average references of the journals, rather than the actual references of each article, is only ∼4%.

5. Conclusion

In conclusion, I have proposed an index of research merit based on creativity, defined as the creation of new scientific knowledge, in a plausible model of knowledge generation and transmission. It is calculated easily from the citations and references of the author's articles, and it is well suited for interdisciplinary comparisons. An advantage of such an index is that its meaning may be more easily perceived, by policy makers and the general public, as a measure of a scientist's social and economic service to the community.

Acknowledgements

I would like to acknowledge very useful discussions with J.V. Alvarez, J.R. Castillo, R. García, J. Gómez-Herrero, L. Seijo, and F. Yndurain. This work has been founded by Spain's Ministry of Science grants BFM2003-03372 and FIS2006-12117.

References

Albert, R., & Barabasi, A. L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics*, *74*, 47–97.

Chen, P., Xie, H., Maslov, S., Redner, S. (2006). Finding scientific gems with Google. Preprint arXiv: physics/0604130.

Cole, J., & Cole, S. (1971). Measuring quality of sociological research—Problems in use of science citation index. *American Sociologist*, *6*, 23–29.

Cozzens, S. E. (1989). What do citations count? The retoric 1st model. *Scientometrics*, *15*, 437–447.

Garfield, E. (1964). Science citation index-new dimension in indexing—Unique approach underlies versatile bibliographic systems for communicating + evaluating information. *Science*, *144*, 649–654.

Gilbert, G. N. (1977). Referencing as persuasion. *Social Studies of Science*, *7*, 113–122.

Gradshteyn, I. S., & Rydhik, I. M. (1980). *Table of integrals, series, and products*. Orlando: Academic Press.

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

Institute for Scientific Information-Thomson Scientific (2006). Webpage: http://isiknowledge.com.

Merton, R. K. (1968). The Matthew effect in science. *Science*, *159*, 56–63.

Moed, H. F. (2005). *Citation analysis in research evaluation*. Dordrecht: Springer.

Narin, F., Hamilton, K. S., & Olivastro, D. (1997). The increasing linkage between US technology and public science. *Research Policy*, *26*, 317–330.

Pinski, G., & Narin, F. (1976). Citation influence for journal aggregates of scientific publications—Theory, with application to literature of physics. *Information Processing and Management*, *12*, 297–312.

Redner, S. (2005). Citation statistics from 110 years of physical review. *Physics Today*, *58*(6), 49–54.

Rinia, E. J., van Leeuwen, T. N., Bruins, E. E. W., van Vuren, H. G., & van Raan, A. F. J. (2002). Measuring knowledge transfer between fields of science. *Scientometrics*, *54*, 347–362.

Rinia, E. J., van Leeuwen, T. N., van Vuren, H. G., & van Raan, A. F. J. (1998). Comparative analysis of a set of bibliometric indicators and central peer review criteria—Evaluation of condensed matter physics in The Netherlands. *Research Policy*, *27*, 95–107.

Soler, J. M. (2006). Webpage: http://www.uam.es/jose.soler/tools.

SPIRES. (2006). Webpage: http://www.slac.stanford.edu/spires/hep/.

Tribus, M. (1969). *Rational descriptions, decisions, and designs*. New York: Pergamon Press.

Available online at www.sciencedirect.com
ScienceDirect
Journal of Informetrics 1 (2007) 161–169
Journal of INFORMETRICS
An International Journal
www.elsevier.com/locate/joi

Comparing alternatives to the *Web of Science* for coverage of the social sciences' literature

Michael Norris, Charles Oppenheim *

Department of Information Science, Loughborough University, Loughborough, Leicestershire LE11 3TU, UK

Received 17 October 2006; received in revised form 5 December 2006; accepted 7 December 2006

Abstract

The *Web of Science* is no longer the only database which offers citation indexing of the social sciences. *Scopus*, *CSA Illumina* and *Google Scholar* are new entrants in this market. The holdings and citation records of these four databases were assessed against two sets of data one drawn from the 2001 Research Assessment Exercise and the other from the *International bibliography of the Social Sciences*. Initially, *CSA Illumina's* coverage at journal title level appeared to be the most comprehensive. But when recall and average citation count was tested at article level and rankings extrapolated by submission frequency to individual journal titles, *Scopus* was ranked first. When issues of functionality, the quality of record processing and depth of coverage are taken into account, *Scopus* and *Web of Science* have a significant advantage over the other two databases. From this analysis, *Scopus* offers the best coverage from amongst these databases and could be used as an alternative to the *Web of Science* as a tool to evaluate the research impact in the social sciences.

Keywords: RAE; *Web of Science*; *Scopus*; *CSA Illumina Google Scholar* research impact

1. Introduction

Since 1986, university research funding decisions in the UK have been based on a series of Research Assessment Exercises (RAE). These have been concerned with assessing the quality of academic research by the peer review of the scholarly output from university departments. Government funding decisions have then been taken based on the ranking of these departments after the review process. This process of assessing academic research by peer review is unlikely to be used after the forthcoming 2008 RAE. It is expected that after this, assessment will be based, in part at least on other means, notably by the use of some form of metrics, possibly including bibliometric measures of research impact. Using such measures relies heavily on having adequate bibliometric records of the subject being assessed. Up until now, this has generally been achieved by using the citation indexes provided by the *Web of Science.* In the case of the social sciences, despite some criticisms of its coverage of the field, the *Web of Science*, until fairly recently was the only credible database which had coverage of the subject and provided citation indexing, thus allowing the possibility of making some measurement of a department's impact in a subject. Recently, however, a number of other providers have entered this market and also offer such a service (Roth, 2005, pp. 1531–1536). *Scopus*, a multidisciplinary database,

* Corresponding author. Tel.: +44 1509 223065; fax: +44 1509 223053.
E-mail addresses: M.Norris2@lboro.ac.uk (M. Norris), C.Oppenheim@lboro.ac.uk (C. Oppenheim).

1751-1577/$ – see front matter © 2007 Charles Oppenheim. Published by Elsevier Ltd. All rights reserved.
doi:10.1016/j.joi.2006.12.001

was launched by Reed Elsevier in 2004 with citation indexing. Likewise, *CSA Illumina* has added this feature to some of its databases and *Google* has added *Google Scholar* to its family of free services as a database of electronic scholarly sources.

It is generally recognised, however, that a good proportion of the scholarly output in the social sciences other than those subjects related to medicine and health are less well covered than the natural sciences in the *Web of Science* (Moed, 2005, pp. 125–126). This lack of coverage is usually attributed to the publication patterns in the social sciences, which to a certain extent favours monographs (Hicks, 2004). Hicks (2004, p. 477) notes that 85% of natural scientists' output is found in journal and conference papers, whilst for social scientists this can range between 42 and 61%. Added to this, there is evidence that many citations in the social sciences are to books rather than journal articles (Hicks, 2004, pp.480–484). Nederhof (2006, p. 83), in his review of research performance in the social sciences, confirms a regional or national orientation to the publications patterns for some of the fields in the social sciences rather than publishing research on an international basis. However, the *Social Science Citation Index* in the *Web of Science* indexes material that is predominantly in English (93–95%) with the remainder being shared between German (2–3%), French (1%) and other languages (Nederhof, 2006, p. 84). There does, however, appear to be a shift towards social scientists publishing more of their work in journals. Larivière et al. (2006, pp. 1002–1003) compared the referencing practices of those in the natural science and engineering fields with those in the humanities and the social sciences and concluded that in the social sciences, there appears to be a steady increase in the share of the number of citations to journals.

Assessing which database to use to measure the scholarly output and impact of the social sciences is an issue of growing importance, given that the *Web of Science* is now not the only choice and that the results of such measurements are likely to assume greater importance in the apportionment of research funding. Ideally, any database which offers coverage of the social sciences would incorporate to a greater degree the scholarly output found in monographs, reports, articles and articles appearing in non-English language journals.

Now that it is possible to assess the coverage of the social sciences by evaluating other databases it seems appropriate to make some assessment of their coverage and quality compared to the *Web of Science*. Usefully, the journal article submissions made to the last RAE in 2001 across 13 Units of Assessment (UoA), which broadly cover the social sciences in the UK, can be used as a benchmarking tool to assess the coverage of these databases (see Appendix A). In particular, using these article submissions, an assessment can be made of the depth of coverage of the:

- Journals in which UK social scientists publish.
- A number of European foreign language journals in the social sciences, notably German, French, Italian and Spanish.

Finding credible sources which could be used to benchmark the coverage of monographs or reports is difficult and hence these have been excluded from the analysis.

2. Database selection

2.1. CSA Illumina

CSA Illumina, formerly *Cambridge Scientific Abstracts*, is a collection of about a 100 bibliographic databases, some of which it hosts with other database partners. The social sciences have a distinct group of databases which includes, for example, *Social Services Abstracts*, *Sociological Abstracts*, and *PsycINFO*. *Sociological Abstracts* is available from 1952, and indexes about 1800 serials, a number of conference proceedings and books. A selection policy classifies journal titles and their contents into one of three categories—'core', 'priority' or 'selective'. Core journals are those journals which are key publications in the discipline and almost all articles are indexed, priority journals are related to the discipline and over 50% of the articles are indexed. Journals which are classed as selective have fewer than 50% of their articles are indexed. Within these databases citation indexing is currently only available in some of them and sometimes this is limited to only core journals. For example, all journal articles contained within the *Social Services Abstracts* have had their cited references indexed from 2004, whilst *Sociological Abstracts* has cited references from core journals only.

2.2. Google Scholar

Google Scholar, launched in its beta form in 2004, provides at no cost multidisciplinary access to scholarly information. It is not clear from which sources *Google Scholar* has built its database, or how large it is, but it is evident that a number of publishers have allowed their electronic journal records to be indexed by them. *Google Scholar* offers basic and (fairly crude) advanced searches, but it is not possible to email, save or manipulate these records in any meaningful way. Citation indexing is available and results are presented roughly in order of the number of times they have been cited.

2.3. Scopus

Scopus, launched in 2004, is a multidisciplinary database with citation indexing. It indexes about 14,000 journal titles, of which about 2850 are from the social sciences (Scopus Content Coverage, 2006, p. 8). Content coverage varies dependant on subject, but for the social sciences this goes back to 1996. From this date, *Scopus* has cover-to-cover indexing of contents, subject to minor exclusions. *Scopus* has basic, author and advanced searches with tools available to manipulate the search results. Cited references can be counted, followed and tracked in the conventional manner with links across the whole database.

2.4. Web of Science

The *Web of Science*, a multidisciplinary database is made up of three citation indexes: *Science Citation Index Expanded*, *Social Sciences Citation Index* and *Arts & Humanities Citation Index*. *Social Sciences Citation Index* started in 1973 and has retrospective coverage going back to 1956. This database indexes the contents of about 1900 journals on a cover-to-cover basis, but it also indexes selectively from over 3000 other journal titles. Additionally, a limited number of conference proceedings and monographs are indexed. The *Web of Science* has general, cited reference and advanced search features, with an extensive range of tools with which to manipulate search results. Cited references can be followed, tracked, counted, processed and analysed across all three of the databases.

2.5. Other source issues

To benchmark foreign language holdings, the *International Bibliography of the Social Sciences*, a bibliography managed by the London School of Economics and Political Science was used. It regularly indexes over 2800 journals which broadly cover the social sciences. A substantial part of these holdings are published in German, French, Italian and Spanish.

3. Methods

The submission records for the 13 UoAs these were extracted from the Higher Education & Research Opportunities (HERO) website which holds the submission records for the last (2001) RAE (HERO, 2006). The 2800 journal titles contained within the IBSS were also extracted (About IBSS, 2006). With the exception of *Google Scholar,* each of the selected databases' journal holdings were also extracted by their respective International Standard Serial Number (ISSN).

3.1. Record processing

Journal articles (which comprised 66.2% of the submissions to the selected UoAs) were checked for their accuracy using their ISSN and journal title against Ulrich's Periodicals Directory (2006). In this process, 720 (2.1%) unverifiable article records were found and discarded. Journal titles were weighted by the frequency with which articles had been submitted to them. Thus, a journal which had one submission to it would have a weighting of one, two submissions would have a weighting of two and so on. Where journals had made clear unambiguous title changes after 2001 and had three or more articles submitted to them in their earlier title, the new journal title and its ISSN were added to the listing with the same journal weighting. Overall 4594 unique journal titles with a total of 33,533 associated article records

were collected. No account was taken of the frequency with which multiple authors of the same article submitted the same article independently of each other to the 2001 RAE.

Turning to the non-English journal titles, these journal records from the IBSS database were checked against Ulrich's Periodicals Directory (2006) to verify their country of origin. This ensured that French titles were in fact published in France rather than, say, Canada or that a Spanish title was published in Spain rather than South America. From this analysis, 581 journal titles were identified as being published in the required countries, split between France 318, Germany 186, Italy 26 and Spain 81.

For the analysis of *Google Scholar's* journal holdings, a statistically valid sample of 380 journal titles, randomly selected, was taken from the complete listing of journal titles submitted to the UoAs. On a similar basis, a sample of 229 records was selected from the 581 IBSS journal titles which cover the four countries given above.

Whilst the selected databases may hold a particular journal title it is important to verify whether they have indexed its content or not. To do this, a statistically valid random sample of 306 articles drawn from different journals was used to check the holdings of the four databases. This process established whether the specific article could be located and the number, if any, of citations to the article.

3.2. Evaluation procedures

The ISSNs from the processed listings were compared record by record to the holdings of *CSA Illumina*, *Scopus* and *WoS*. From this process, a count and percentage coverage was obtained for each database. A weighted holding for each database was calculated based on the matching journal records found and the frequency with which articles had been submitted to them. This, for example, weighted the *British Journal of Sociology* at 73, given that it had had 73 of its articles submitted to the 2001 RAE. Using the same method, the ISSNs from the IBSS foreign journal title listing was compared.

Using *Google Scholar's* advanced search feature, the 380 records selected through the sampling process were entered by journal title. The matching records from these searches were ranked into the following categories:

- Zero hits;
- citation only records;
- multiple websites listing the journal and various articles from them; and
- consistent hits leading to a single credible website.

Zero hits indicate no matching records were found. Citation only records are where matching records to the journal article have been cited but the original article record could not be found. Where multiple websites reported articles from the journal, these derived from publishers, aggregators or open access sources. Hits showing a high degree of uniformity usually led to the publisher's own website.

4. Results

4.1. Overall coverage

Table 1 shows the number of matching records found and the percentage coverage for each database against the 4594 unique journals titles identified in the process described above. The weighted UoA takes the matching journal titles found for each database and gives the sum of the number of articles submitted to them and the percentage reported is of all the articles submitted.

Table 1
Record count and percentage coverage for each database

Database	All UoAs (titles)	Weighted UoA (articles)
Web of Science	1994: 43.4%	20,265: 60.4%
Scopus	2324: 50.6%	22,996: 68.6%
CSA Illumnia	2678: 58.3%	24,436: 72.9%

Table 2
Foreign journal coverage

	Web of Science	*Scopus*	*CSA*
IBSS (581 titles)	71: 12.2%	61: 10.5%	140: 24.1%

Table 3
Coverage by *Google Scholar*

	All UoAs	IBSS
Zero hits	15: 3.9%	211: 92.1%
Citations only	38: 10.0%	3: 1.3%
Multiple websites	111: 29.2%	8: 3.5%
Single website	216: 56.8%	7: 3.1%

Taking the journal holdings from the IBSS for France, Germany, Spain and Italy and comparing this to the holdings of each of the three databases in the same manner as Table 1 above, then coverage details are as given in Table 2.

Of the 581 IBSS journals identified for benchmarking, 92 (15.8%) were used by authors in their submissions to the 2001 RAE.

4.2. Google Scholar

Extrapolated coverage of the 4594 UoA journal titles and the 581 foreign journal titles by *Google Scholar* are given in Table 3 below. These are ranked by the nature of the hits, as defined above. Sample sizes were, respectively, 380 from the 4594 UoA titles, and 229 from the 581 foreign language titles.

The 216 records representing 56.9% of the UoA sample, and the 7 records from IBSS which could be attributed to a single website, can be compared to those matches found by the record by record comparison for the other databases.

4.3. Article and citation coverage

The coverage test at journal, article and citation level showed that all the databases, with the exception of one title in *CSA Illumina,* had the journal title records that they claimed to hold, albeit, at differing levels. *Google Scholar's* holdings are of course unknown so it is not possible to say what may be missing. Table 4 shows the coverage.

Analysis of 15 records from the 65 article records not found by *CSA Illumina* showed that 11 of these had not been indexed, even though other articles from the same year or journal issue had been. In the case of the other four articles, the year in which they were published the journal did not appear to have been indexed at all. Further analysis showed that 14 of these journals had the minimal 'selective' category for indexing their content, and the remaining journal had 'priority' status only.

Using the 'Journal and article found percentage coverage' from Table 4 with the earlier results from Table 1 for the 'Weighted UoA', an overall estimate of the coverage was calculated. These are shown in column 4 in Table 5 below. This result is based upon taking the journal and article found percentage and assuming that this find rate can be used

Table 4
Coverage at article level

	Web of Science	*Scopus*	*CSA Illumina*	*Google Scholar*
Journal not found	0	0	1	12
Citation only record (GS only)	–	–	–	29
Journal found but not the specific article record searched for	37	15	65	0
Journal and article found and percentage coverage	269, 87.9%	291, 95.1%	240, 78.4%	265, 86.6%
Articles found but no citations noted	31	27	140	25
Average citation count per article found	13.7	14.5	3.1	17.7

Table 5
Record count and percentage coverage for each database

Database	All UoAs titles	Weighted UoA by article frequency	Re-weighted by percentage found
Web of Science	1994: 43.4%	20,265: 60.4%	17,782: 53.1%
Scopus	2324: 50.6%	22,996: 68.6%	21,869: 65.2%
CSA	2678: 58.3%	24,436: 72.9%	19,158: 57.1%

Table 6
Paired coverage of submissions to the 2001 RAE

Additions	*CSA Illumina*		*Scopus*		*Web of Science*	
	Scopus	*WoS*	*CSA*	*WoS*	*CSA*	*Scopus*
Titles	497	361	851	170	1045	500
Weighted article count	3423	2415	4863	707	6586	3438
Article recall adjustment	3255	2123	3813	621	5163	3270

Table 7
Combined coverage

	CSA Illumina		*Scopus*		*Web of Science*	
	Scopus	*WoS*	*CSA*	*WoS*	*CSA*	*Scopus*
Total titles	3,175	3,039	3,175	2,494	3,039	2,494
Total weighted articles	27,859	26,851	27,859	23,703	26,851	23,703
Article recall adjustment	22,413	21,281	25,682	22,490	22,945	21,052
Overall article coverage %	66.8	63.5	76.6	67.1	68.4	62.8

as a satisfactory measure of recall rate. Multiplying this recall rate against the original 'Weighted UoA' score gives the expected coverage from the articles submitted to the 2001 RAE.

4.4. Overlap analysis

There is a common set of 1516 journals that each of the three databases share with each other, and these account for 17,253 (51.4%) of the 33,533 articles. Table 6 below shows the overlap of the databases in relation to each other. The relationship shows the effect of combining each of the databases on a paired basis. For example, *CSA Illumina* covers 2678 of the journal titles submitted to the 2001 RAE and is shown as the primary source of coverage. *Scopus* covers a further 497 journal titles that *CSA Illumina* does not cover and their value is shown in terms of the 3423 articles that those additional titles would bring. This additional article count is then adjusted by the percentage coverage factor from Table 4; hence the article coverage is adjusted by a factor 0.951.

Table 7 combines the results of Tables 5 and 6. The original 2678 titles covered by *CSA Illumina* shown in Table 5 have been combined with the 497 additional titles from *Scopus* shown in Table 6 to give a total of 3175 journal titles which the two databases cover from the 2001 RAE submissions. Likewise, the article counts have been combined along with adjustments for the percentage coverage factor. The 'Overall article coverage %' varies between each pairing because each of the databases has a different percentage coverage factor. Finally, an overall coverage figure is given in terms of the number of articles covered compared to the total of 33,533 articles that were identified.

5. Discussion

5.1. Coverage

From the initial analysis (see Tables 1 and 2), it appears that *CSA Illumina* has the greater coverage in terms of journal titles for all the UoAs taken together and the IBSS database collection. In a simple ranking by total jour-

nal coverage, *CSA Illumina* is followed by *Google Scholar*, *Scopus* and then by *Web of Science*. *Google Scholar* was, however, the least successful database in the way that it presented the results it did find. Whilst the other databases presented their results chronologically, in an orderly manner, by volume and then by issue, the results from *Google Scholar* were generally ranked by citation count irrespective of volume or issue sequence. *CSA Illumina* provided the best coverage of the foreign journals selected from the IBSS database, by a factor of almost 2 when compared to the *Web of Science* with *Google Scholar* giving the worst coverage, finding only 7 of 229 titles tested.

Whilst *CSA Illumina*'*s* success at journal title level appears to be good, this good result was not maintained when individual articles submitted to the 2001 RAE were presented to it. From the sample of 306 articles presented to each of the databases, *CSA Illumina* was the least successful, scoring only 240 hits out of the 306 possible. *Scopus* was best at 291 hits with *Web of Science* finding 269 and *Google Scholar* finding 265. The *Web of Science* and *Scopus* appear to have a broader and more robust 'cover to cover' indexing policy, *CSA Illumina*'*s* policy appears to be more variable and is dependent on how close the journal's contents is to the main discipline of the database in which it is held. Examination of the selection policy of the 15 sample journals where articles could not be found showed that their selection was not 'core', so hence not all articles would have been indexed.

Given the newness of *Scopus* and *Google Scholar,* these two databases have been frequently reviewed and compared and in several cases they have been compared to the *Web of Science.* In a number of papers, Jacso (2005a, pp. 208–214, 2005b, pp. 1537–1547) has discussed the limitations of *Google Scholar*. He has concluded that it is unreliable and unpredictable in the results it returns, both in its links to the sources it has found and in its coverage. This view of *Google Scholar* is also shared, generally, by others who have also found significant omissions in the coverage and recall from this database (Myhill, 2005; Notess, 2005). It is evident, however, that most reviewers feel that *Google Scholar* has the potential to become a useful source of scholarly information provided its shortcomings are addressed.

Like *Google Scholar*, *Scopus* has been subject to close scrutiny, Deis and Goodman (2005) have compared the *Web of Science* and *Scopus* as has LaGuardia (2005). Others (Burnham, 2006; Dess, 2006; Jacso, 2004) have also reviewed *Scopus* extensively. These reviewers generally acknowledge that both of these databases are primarily concerned with the sciences. The number of science based records held by each greatly outweighs their holdings in the social sciences and the arts. Nevertheless, both databases still have large holdings of records in the social sciences and these are comprehensively indexed, although *Scopus* has only done this from 1996. This lack of coverage by *Scopus* is a noted concern, Dess (2006) recommends the *Web of Science* as the only plausible option for searching prior to 1996 for interdisciplinary subjects such as the social sciences. When, however, Dess (2006) carried out citation searches for post 1996 records, *Scopus* had a small advantage of 1.2% in the number of citations that it found over the *Web of Science*. The difference found in the citation count test here, in average citations per article found was a 5.4% advantage in favour of *Scopus*.

5.2. Citation coverage

The use of citation counts to measure research performance in the social sciences and humanities is more problematic than in the sciences. Nederhof (2006, pp. 81–100) acknowledges the shortcomings of the *Web of Science* to cover adequately the social sciences and in particular the humanities, given the propensity of academics in the latter to publish more in monographs than in serials.

Despite this tendency to publish in non-journal sources in the humanities, Nederhof (2006, p. 86) in his review considers that "In most disciplines in the social sciences and humanities, journals were found to be the single most important medium for publication. [and] In the behavioural sciences and economics, journal articles account for the majority of citations". There is a recognised trend: by social scientists publish more of their work in journals than was formerly the case. Adams (n.d.) has analysed the increase in journal submissions between the 1996 and 2001 RAE, and found that these submissions increased as a percentage of all submissions from 42.4% in 1996 to 54.1% in 2001 for the social sciences. In the UoAs included for analysis in this work, journal articles made up 66.2% of all submissions.

The citation count results from Table 4 show that there is significant variation in the number of citations that have been indexed between the different databases. In many cases, individual article counts from the *Web of Science* and *Scopus* were very closely matched. *CSA Illumina* reported the lowest citation counts and had the greatest number

of articles without any citations. This is not surprising since where *CSA Illumina* does index cited references, it has done so only fairly recently and for a limited range of its databases (Cited Reference Linking, n.d.). It also has very limited citation links between citing records from different, but related databases. Whilst *Google Scholar* has the highest citation counts, when these are examined individually, it is clear that the results have not, at the very least, been de-duplicated. This makes the citation counts from *Google Scholar* highly suspect. Jacso (2006) in an examination of how *Google Scholar* counts and reports citation counts, found that it was consistently unreliable.

Clearly, citation coverage varies, depending on the database being used. *Web of Science* has the greater coverage historically and its functionality and sophistication exceeds the other databases considered. *Scopus* has the second largest coverage, in terms of citations and the closest functionality to that offered by *Web of Science*. Whilst the depth of cited reference coverage is not as great as *Web of Science*, cited references in *Scopus* can be searched, analysed and processed in a very similar way, although this may take a little longer than for the *Web of Science*. Citation indexing in *CSA Illumina* is currently only available in some of the databases and is often limited to only core journals. All of the databases other than *Google Scholar* returned search results by volume and issue number or in an order selected by the user.

6. Conclusions and recommendations

All of the databases examined give significant coverage of the social sciences in terms of journals and articles indexed as well as the indexing of cited references. There are, however, noticeable differences between the databases. In its current form, *Google Scholar* cannot seriously be thought of as a database from which metrics could be used to measure scholarly activity.

Whilst *CSA Illumina* has the greatest journal coverage overall, its coverage at article level is disappointing as is the number of journals from which it indexes cited references. However, when the results are weighted by article submission frequency, *CSA Illumina* does rank second in its coverage. This ranking is improved to first position when the database has its coverage combined with *Scopus* or the *Web of Science* as major partners; this is, respectively, 76.6 and 68.4%. Despite this good coverage in combination, when compared to the other databases, however, the citation count from *CSA Illumina* per article appears to be significantly out of step, suggesting that the base from which it is collecting these records is too small. The analysis of cited references is limited, users can identify citing and cited authors; but analysis and record processing are less sophisticated than *Scopus* or the *Web of Science. CSA Illumina* falls too far below the standards set by *Web of Science* and *Scopus*, even though it has good article coverage, to be considered a serious option to cover the social science literature and in particular, to be used to measure scholarly activity and impact in this field.

Taking an overall view, both *Web of Science* and *Scopus* offer the best coverage at journal, article and cited reference level. Both index their journal holdings on a cover to cover basis. They would together give an article coverage rate of 67.1%. Both, on the other hand, are rather weak on their coverage of foreign journals and *Scopus* does not currently go further back than 1996 for the social sciences. Given that coverage by *Scopus* is considered good, and its tools to analyse citation counts are sufficient, then arguably it offers the best choice from amongst the multidisciplinary databases reviewed here. On this basis, notwithstanding its poor foreign journal title coverage, it is suggested that *Scopus* could be used as an alternative to the *WoS* as a tool to evaluate research impact in the social sciences.

As indicated earlier, finding an adequate set of book and report records with which to credibly benchmark database holdings from the 2001 RAE submissions has not been successful. Examination of the submissions to the 2001 RAE show that reports make up less than 2% of the total and Nederhof (2006, p. 95) in his review suggests that for ‘grey’ publications such as unpublished reports, that their impact is “rather disappointing”. Although monographs remain an important social sciences scholarly output, recent research has shown that there is a steady movement towards publication in journals and away from monographs. It is expected that this trend will continue. Therefore, the significance of this lack of coverage will decline over time.

Acknowledgements

We are grateful for the ESRC for funding this research and to Dr. J. Cooper and R. Cornish for technical help and statistical advice.

Appendix A

Units of Assessment (UoA)

UoA number	Name
13	Psychology
34	Town and Country Planning
35	Geography
37	Anthropology
38	Economics and Econometrics
39	Politics and International Studies
40	Social Policy and Administration
41	Social Work
42	Sociology
43	Business and Management Studies
44	Accounting and Finance
56	Linguistics
68	Education

References

About IBSS. (2006). <http://www.lse.ac.uk/collections/IBSS/about/alphabeticalJournals.htm> Accessed 15.05.06.

Adams, J. (n.d.). *Research Assessment and UK publication patterns*. <www.uksg.org/presentations8/adams.pps> Accessed 22.05.06.

Burnham, J. (2006). Scopus database: A review. *Biomedical Digital Libraries*, *3*(1). <http://www.bio-diglib.com/content/3/1/1> Accessed 30.05.06.

Cited Reference Linking. (n.d.). <http://www.csa.com/help/Advanced_Search/cited_reference.html> Accessed 05.05.06.

Deis, L. & Goodman, D. (2005). Web of Science (2004 version) and Scopus. *The Charleston Advisor*, *6*(3). <http://www.charlestonco.com/comp.cfm?id=43> Accessed 03.05.06.

Dess, H. (2006). Database reviews and reports: Scopus. *Issues in Science and Technology Librarianship*. <http://www.istl.org/06-winter/databases4.html> Accessed 03.05.06.

HERO Higher Education Research Opportunities. (2006). <http://www.hero.ac.uk/rae/Results/index.htm> Accessed 08.05.06.

Hicks, D. (2004). The four literatures of the social science. In H. K. Moed, W. Glanzel, & U. Schmoch (Eds.), *Handbook of Quantitative Science and Technology Research* (pp. 473–495). Dordrecht: Springer.

Jacso, P. (2004). *Péter's Digital Reference Shelf. Scopus.* <http://www.galegroup.com/reference/archive/200409/scopus.html> Accessed 20.09.05.

Jacso, P. (2005a). Google Scholar: The pros and the cons. *Online Information Review*, *29*(2), 208–214.

Jacso, P. (2005b). As we may search—Comparison of major features of *Web of Science*, *Scopus* and *Google Scholar* citation-based and citation-enhanced databases. *Current Science*, *89*(9), 1537–1547.

Jacso, P. (2006). Deflated, inflated and phantom citation counts. *Online Information Review*, *30*(3), 297–330.

LaGuardia, C. (2005). *E-Views and Reviews: Scopus vs Web of Science. Library Journal.com*. <http://www.libraryjournal.com/article/CA491154.html%22> Accessed 03.05.06.

Larivière, V., et al. (2006). The place of serials in referencing practices: Comparing natural sciences and engineering with social sciences and humanities. *Journal of the American Society for Information Science and Technology*, *57*(8), 997–1004.

Moed, H. K. (2005). *Citation analysis in research evaluation*. Dordrecht: Springer. Chapter 7: ISI Coverage by discipline

Myhill, M. (2005). *Google Scholar*. <http://www.charlestonco.com/review.cfm?id=225> Accessed 03.05.06.

Nederhof, A. (2006). Bibliometric monitoring of research performance in the social sciences and the humanities: A review. *Scientometrics*, *66*(1), 81–100.

Notess, G. (2005). *Scholarly web searching: Google Scholar and Scirus*. <http://www.infotoday.com/Online/jul05/OnTheNet.shtml> Accessed 03.05.06.

Roth, D. (2005). The emergence of competitors to the *Science Citation Index* and the *Web of Science*. *Current Science*, *89*(9), 1531–1536.

Scopus Content Coverage. (2006). <http://www.info.scopus.com/docs/content_coverage.pdf> Accessed 19.06.06.

Ulrich's Periodicals Directory. (2006). <http://www.ulrichsweb.com/ulrichsweb/> Accessed 08.05.06.

Journal of Informetrics

Aims and Scope

Journal of Informetrics (JOI) publishes refereed articles on fundamental quantitative aspects of information science. The journal, although limited to -metrics aspects, has a broad scope: in principle, all quantitative analysis of original problems in information science are within the scope of JOI. Besides its generality, Journal of Informetrics focusses on papers describing fundamental methods and theories and/or universally important data, gathered in a non-trivial way. Fundamental methods comprise mathematical, probabilistic or statistical models and techniques as well as methods in operational research. These methods can serve the quantitative explanation of certain phenomena, evaluation of information and its producers as well as the management of libraries and other information centres.

Journal of Informetrics has a special (though not exclusive) interest in inter- and multi-disciplinary papers, dealing with common aspects of (or possible differences between) several neighbouring disciplines such as quantitative linguistics, econometrics, biometrics and other -metrics fields. The aim is to lower the barriers between these fields, hence avoiding reformulation of similar problems, theories and solutions. Journal of Informetrics also welcomes certain papers from researchers who do not consider themselves as informetrists, for example research papers would be considered on the graph-theoretic description of networks.

Journal of Informetrics also publishes papers that improve standardisation in informetrics. In general the journal aims to contribute to increasing the degree of "hardness" of the field, and to increase the degree of "exactness" of the scientific field of informetrics.

The journal covers informetrics and considers it to comprise (or at least to include) fields such as bibliometrics, scientometrics, webometrics and cybermetrics. Specific topics can be described (non-exhaustively) as follows: informetric laws (including, but not exclusively: Lotka, Zipf, Bradford, Mandelbrot but also laws of growth and ageing or obsolescence) hereby also modelling generalised bibliographies, aspects of inequality or concentration (e.g. Lorenz theory) and diffusion, citation theory, linking theory, downloads, indicators (definitions and properties), evaluation techniques for scientific output (literature, persons) and for documentary systems (information retrieval) incl. ranking theory, library management, graph-theoretic and topological analysis of networks (incl. Internet, intranets, citation and collaboration networks), visualisation and mapping of science (persons, fields, institutes, topics,...).

A full Guide for Authors can be found in INFORMETRICS 1/1, or online on the journal website at: www.elsevier.com/locate/joi.

Available online at www.sciencedirect.com
ScienceDirect
Journal of Informetrics 1 (2007) 115–122
Journal of INFORMETRICS
An International Journal
www.elsevier.com/locate/joi

General evolutionary theory of information production processes and applications to the evolution of networks

L. Egghe [a,b]

[a] *Universiteit Hasselt (UHasselt), Campus Diepenbeek, Agoralaan, B-3590 Diepenbeek, Belgium*
[b] *Universiteit Antwerpen (UA), Campus Drie Eiken, Universiteitsplein 1, B-2610 Wilrijk, Belgium*

Received 28 August 2006; received in revised form 2 October 2006; accepted 11 October 2006

Abstract

Evolution of information production processes (IPPs) can be described by a general transformation function for the sources and for the items. It generalises the Fellman–Jakobsson transformation which only works on the items.

In this paper the dual informetric theory of this double transformation, defined by the rank-frequency function, is described by, e.g. determining the new size-frequency function. The special case of power law transformations is studied thereby showing that a Lotkaian system is transformed into another Lotkaian system, described by a new Lotka exponent. We prove that the new exponent is smaller (larger) than the original one if and only if the change in the sources is smaller (larger) than that of the items.

Applications to the study of the evolution of networks are given, including cases of deletion of nodes and/or links but also applications to other fields are given.

Keywords: Evolution; IPP; Information production process; Lotka; Zipf; Network

1. Introduction

The informetrics of information production processes (IPPs) can be described via the so-called size-frequency function f:

$$f : [a, \rho_m] \to \mathbb{R}^+; \quad j \to f(j) \tag{1}$$

where $f(j)$ denotes the density of the sources in item density j: this is the continuous extension for the classical $f(j)$ = number of sources with j items and we let $j \geq a \geq 1$ also be limited to a maximal item density ρ_m (see also Egghe, 2005 but there $a = 1$; here we use a general $a > 0$ since we have an application of this case—see further). A classical example is the law of Lotka, where f is then a decreasing power law; this case will also be considered after the general theory.

The size-frequency function f is equivalent with the rank-frequency function g:

$$g : [0, T] \to \mathbb{R}^+; \quad r \to g(r) \tag{2}$$

E-mail address: leo.egghe@uhasselt.be.

1751-1577/$ – see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.joi.2006.10.003

where f and g are related as

$$r = g^{-1}(j) = \int_j^{\rho_m} f(k)\,dk \tag{3}$$

where g^{-1} denotes the inverse function of g. It is clear from (3) that $g(r)$ denotes the item density in the source on rank density r: this is the continuous extension of the discrete rank-frequency function where $g(r)$ = number of items in the source on rank r and where T denotes (also in the continuous setting) the total number of sources.

The equivalence of the functions f and g is seen as follows: (3) yields g^{-1} (hence g), given f and it follows from (3) that

$$f(j) = -\frac{1}{g'(g^{-1}(j))} \tag{4}$$

for all $j \in [a, \rho_m]$, hence f follows from g, showing the equivalency (see also Egghe, 2005). It is also well-known (see Egghe, 2005 or Egghe and Rousseau, 1990) that, in case f is a decreasing power law (i.e. Lotka's law), g is the so-called law of Mandelbrot (which we will describe in detail below).

In Egghe (2004), see also Egghe (2003) one studies positive reinforcement of IPPs, where one applies a transformation φ on the function g, i.e. g is transformed into $g^* = \varphi^\circ g$, where φ has certain properties, e.g. $\varphi(x) \geq x$ for all x and φ strictly increasing. In Egghe (2003, 2004), the connection of positively reinforced IPPs with linear 3-dimensional informetrics (i.e. the composition of 2 IPPs) is highlighted and the concentration properties of these positively reinforced IPPs are indicated using the theorem of Fellman and Jakobsson (see Fellman, 1976; Jakobsson, 1976; see also Egghe, 2007).

In Egghe (2003, 2004) and Egghe and Rousseau (2006a), a transformation of g in the following sense has been studied:

$$g^*(r) = B(g(r))^c \tag{5}$$

with B, $c > 1$ (i.e. $\varphi(x) = Bx^c$) yielding, for Lotkaian IPPs, lower Lotka exponents. In Egghe and Rousseau (2006a) the extra generalization $j \in [a,\rho_m]$ with $a \geq 1$ is used. In this case the transformation (5) not only leads to lower Lotka exponents but also to higher minimum density values $a > 1$. This, in turn, gives a rationale for systems in which sources do not have a low number of items as is the case for database sizes or country or city sizes. That these cases go together with low values of the exponent of the Lotka function has been experimentally verified in Egghe and Rousseau (2006a).

Discussions with Cothey (July 2005) revealed that an extra generalization of the above formalism (essentially the transformation $g \to \varphi^\circ g$) is needed. Indeed, the transformation φ is a transformation that applies on the item densities $j = g(r)$ but leaves the source rank densities unchanged. Cothey informed us that the framework of IPPs is well applied to networks (where sources are nodes and items are hyperlinks: in- or outlinks) but that a model is needed, e.g. to describe disappearing sources (nodes)—of course still allowing for disappearing items as well. Of course the creation of sources and items should also be covered.

In view of the above it is clear what to do: the transformation φ above, in its full generality, works well to describe changes (dynamics) of items. So "all we have to do" is to introduce another transformation, called ψ below, in order to describe the changes (dynamics) of the sources.

In the next section the second transformation ψ will act on the rank densities r. So, instead of the transformation

$$g^*(r) = \varphi(g(r)) \tag{6}$$

we will generalise (6) as follows:

$$g^*(r^*) = g^*(\psi(r)) = \varphi(g(r)) \tag{7}$$

so that also the source rankings are transformed. This very general model (7) will be studied in the next section and its equivalent size-frequency function f^* will be calculated.

In Section 3, the results obtained will be applied to Lotkaian systems and to transformations φ and ψ of power law type. Also in this case the equivalent size-frequency function f^* will be calculated thereby extending the results in Egghe (2003, 2004) and Egghe and Rousseau (2006a).

Several applications of these results are described in Section 4. The applications go from general IPPs to countries or city size distributions, database distributions or (as initiated by Cothey) network distributions and their dynamics (evolutions).

2. General evolutionary model for IPPs

Let us have a first system (IPP) given by $f : [a, \rho_{\mathrm{m}}] \to \mathbb{R}^+$, $j \to f(j)$, as size-frequency function and by its equivalent (cf. (3), (4)) rank-frequency function $g : [0, T] \to \mathbb{R}^+$, $r \to g(r)$. Suppose this system is "changing" into a new system that we describe by asterisks: $f^* : [a^*, \rho^*_{\mathrm{m}}] \to \mathbb{R}^+$, $j^* \to f^*(j^*)$ and $g^* : [0, T^*] \to \mathbb{R}^+$, $r^* \to g^*(r^*)$.

To allow for the largest possible freedom of evolution of the first IPP into the second we allow for a transformation of the source densities as well as of the item densities as follows: we define

$$g^*(r^*) = g^*(\psi(r)) = \varphi(g(r)) \tag{8}$$

where

$$\psi : [0, T] \to [0, T^*]; \quad r \to r^* = \psi(r) \tag{9}$$

is differentiable and where

$$\varphi : [a, \rho_m] \to [a^*, \rho^*_m]; \quad j \to j^* = \varphi(j) \tag{10}$$

is differentiable.

Formula (8) describes the general rank-frequency transformation $g \to g^*$. The corresponding size-frequency transformation $f \to f^*$ is given by the next basic theorem.

Theorem 2.1. *Formula (8) implies, for all $j \in [a, \rho_{\mathrm{m}}]$ and $j^* \in [a^*, \rho^*_m]$:*

$$f^*(j^*) = f(j)\frac{\psi'(g^{-1}(j))}{\varphi'(j)} \tag{11}$$

where $j^* = \varphi(j)$ as above (and assuming $\varphi' \neq 0$).

Proof. By the defining relation (3) we have

$$r = g^{-1}(j) = \int_j^{\rho_m} f(k)\,\mathrm{d}k \tag{12}$$

for all $j \in [a,\rho_{\mathrm{m}}]$, $r \in [0,T]$ and

$$r^* = g^{*-1}(j^*) = \int_{j^*}^{\rho^*_m} f^*(k^*)\,\mathrm{d}k^* \tag{13}$$

for all $j^* \in [a^*, \rho^*_{\mathrm{m}}]$, $r^* \in [0,T^*]$.

Hence, by (9), we have

$$\psi\left(\int_j^{\rho_{\mathrm{m}}} f(k)\,\mathrm{d}k\right) = \int_{j^*}^{\rho^*_{\mathrm{m}}} f^*(k^*)\,\mathrm{d}k^* \tag{14}$$

So, by (10)

$$\psi\left(\int_j^{\rho_{\mathrm{m}}} f(k)\,\mathrm{d}k\right) = \int_{\varphi(j)}^{\rho^*_{\mathrm{m}}} f^*(k^*)\,\mathrm{d}k^* \tag{15}$$

Differentiating both sides of (15) with respect to j yields:

$$\psi'\left(\int_j^{\rho_{\mathrm{m}}} f(k)\,\mathrm{d}k\right)(-f(j)) = -f^*(\varphi(j))\varphi'(j)$$

hence, by (10):

$$f^*(j^*) = f(j)\frac{\psi'\left(\int_j^{\rho_{\mathrm{m}}} f(k)\,\mathrm{d}k\right)}{\varphi'(j)}$$

which gives (11) by (12). □

Corollary 2.1. *If* $\psi = Id$ (i.e. $\psi(r) = r$ *for all* $r \in [0,T]$) *we have, for all* j *and* j^* *as in* Theorem 2.1

$$f^*(j^*) = \frac{f(j)}{\varphi'(j)} \tag{16}$$

Hence,

$$f^*(j^*) = \frac{f(\varphi^{-1}(j^*))}{\varphi'(\varphi^{-1}(j^*))} \tag{17}$$

Proof. This is trivial since $\psi' = 1$ and by (10). □

This special case was already recovered in Egghe (2003, 2004) (for $a = a^* = 1$).

3. Power law transformations in Lotkaian IPPs

Now the obtained results will be applied to Lotkaian IPPs where φ and ψ are transformations of power law type. Lotkaian IPPs are IPPs where we have a decreasing power law for the size-frequency function f:

$$f(j) = \frac{C}{j^\alpha} \tag{18}$$

$C > 0$ and $\alpha > 1$ constants, $j \in [a, \rho_m]$.

Case 3.1 *(*$\rho_m < \infty$*).* As proved in Egghe and Rousseau (1990) – see also Egghe (2005) – we now have for the rank-frequency function g (equivalent to f in (18)):

$$j = g(r) = \frac{E}{(1 + Fr)^\beta} \tag{19}$$

with

$$\beta = \frac{1}{\alpha - 1} \tag{20}$$

$$E = \rho_m \tag{21}$$

$$F = \frac{\alpha - 1}{C\rho_m^{1-\alpha}} \tag{22}$$

Note that the value of a is only implicitly involved in (19), being the lowest possible value for $g(r)$ (i.e. $a = g(T)$).

In this Lotkaian framework, we will also use (increasing) transformations φ and ψ of power type:

$$r^* = \psi(r) = Ar^b \quad (r \in [0, T]) \tag{23}$$

$$j^* = \varphi(j) = Bj^c \quad (j \in [a, \rho_m]) \tag{24}$$

with $A, B, b, c > 0$. We now have the transformation

$$g^*(r^*) = g^*(Ar^b) = B(g(r))^c \tag{25}$$

as follows from (8).

We will now evaluate the form of the transformed size-frequency function f^*. By (11) we have, since

$$\psi'(r) = Abr^{b-1} \tag{26}$$

$$\varphi'(j) = Bcj^{c-1} \tag{27}$$

that

$$f^*(j^*) = f(j)\frac{Ab(g^{-1}(j))^{b-1}}{Bcj^{c-1}} \tag{28}$$

Since f is Lotkaian (18) we have (19) and (20) hence

$$r = g^{-1}(j) = \frac{(E/j)^{\alpha-1} - 1}{F} \tag{29}$$

Substituting (29) in (28) yields

$$f^*(j^*) = \frac{C}{j^\alpha} \frac{Ab[(E/j)^{\alpha-1} - 1]^{b-1}}{F^{b-1}Bcj^{c-1}}$$

$$f^*(j^*) = \frac{CAb}{F^{b-1}Bc} \frac{1}{j^{\alpha+c-1}} \left[\left(\frac{E}{j}\right)^{\alpha-1} - 1\right]^{b-1} \tag{30}$$

Now use (24) yielding

$$j = \left(\frac{j^*}{B}\right)^{1/c} \tag{31}$$

Formula (31) in (30) yields

$$f^*(j^*) = \frac{CAbB^{(\alpha+c-1)/c}}{F^{b-1}Bc} \frac{1}{j^{*(\alpha+c-1)/c}} \left[\frac{E^{\alpha-1}B^{(\alpha-1)/c}}{j^{*(\alpha-1)/c}} - 1\right]^{b-1}$$

Note that the expression between [], by (29) and (31) equals Fr.

So, for $r \in [0,T]$ large enough we have (since $Fr \approx Fr + 1$)

$$f^*(j^*) \approx \frac{CAbB^{(\alpha-1)/c}E^{(\alpha-1)(b-1)}B^{((\alpha-1)(b-1))/c}}{F^{b-1}c} \frac{1}{j^{*\delta}} \tag{32}$$

with

$$\delta = \frac{\alpha + c - 1 + (\alpha - 1)(b - 1)}{c} \tag{33}$$

and for $j^* \geq \varphi(a) = Ba^c$ (since $j \geq a$). Hence, denoting the intricate constant before $1/j^{*\delta}$ in (32) by G, we have, for large r (and for all r if $b = 1$)

$$f^*(j^*) \approx \frac{G}{j^{*\delta}} \tag{34}$$

with δ as in (33), i.e. Lotka's law with exponent δ. Note that (34) is an equality if $b = 1$. So we have proved the following theorem.

Theorem 3.1. *Let φ, ψ and f be as in (18), (23) and (24). Then the transformed size-frequency function f^* has the form*

$$f^*(j^*) \approx \frac{G}{j^{*\delta}}$$

with δ as in (33), i.e. a Lotkaian size-frequency function where the exponent δ is the function (33) of the exponents b, c of the transformations ψ and φ, respectively, and of α, the Lotka exponent of f and where $j^* \geq \varphi(a) = Ba^c$. If $b = 1$ then $\approx$ in (34) is an exact equality.

Case 3.2 *($\rho_m = \infty$).* Next we prove the same result for $\rho_m = \infty$. We can now prove that (34) holds with an exact equality:

Theorem 3.2. *Let φ, ψ and f be as in (18), (23) and (24) with $\rho_m = \infty$. Then the transformed size-frequency function f^* has the form*

$$f^*(j^*) = \frac{G}{j^{*\delta}} \tag{35}$$

with δ as in (33) and $j^* \geq \varphi(a) = Ba^c$.

Proof. Instead of (19) we now have—see Egghe (2005), Exercise II.2.2.6 or Egghe and Rousseau (2006b) (Appendix) where a proof is provided:

$$g(r) = \frac{E}{r^{\beta}} \tag{36}$$

with $r \in [0,T]$, $E>0$ and β as in (20). Since now

$$g^{-1}(j) = \left(\frac{E}{j}\right)^{1/\beta} \tag{37}$$

it follows from (11), (26), (27) and (37) that

$$f^*(j^*) = \frac{CAb(E/j)^{(b-1)/\beta}}{j^{\alpha}Bcj^{c-1}} \tag{38}$$

Now (31) yields, using (20)

$$f^*(j^*) = \frac{\mathrm{CAbE}^{(b-1)(\alpha-1)}B^{((\alpha-1)/c)b}}{cj^{*\delta}} \tag{39}$$

with δ as in (33), exactly and where $j^* \geq \varphi(a) = Ba^c$. □

We have the following trivial but important proposition.

Proposition 3.2. *In the notation of above, we have*

(i) $\delta<\alpha \Leftrightarrow b<c$
(ii) $\delta=\alpha \Leftrightarrow b=c$
(iii) $\delta>\alpha \Leftrightarrow b>c$

Proof. We only prove (i); the proof of (ii) and (iii) is similar. By (33):

$$\delta = \frac{\alpha + c - 1 + (\alpha - 1)(b - 1)}{c} < \alpha$$

iff

$$\alpha + c - 1 + (\alpha - 1)(b - 1) < \alpha c$$

iff

$$(\alpha - 1)(b - 1) < (\alpha - 1)(c - 1)$$

iff

$$b < c$$

since $\alpha>1$. □

The interpretation of this corollary is important: Corollary 3.2 gives necessary and sufficient conditions for the evolution of Lotkaian IPPs to result in higher or lower (or constant) Lotka exponents. In terms of sources and items this means, by (23) and (24), that Lotka's exponent δ is decreasing under the transformation (i.e. $\delta<\alpha$) *if and only if* the "change" in the sources is smaller than the one in the items ($b<c$). Analogous for the other assertions. Note also that if the transformations φ and ψ have the same exponents then $\delta=\alpha$, hence Lotka's exponent remains the same.

Note also that $\delta = (c + (\alpha - 1)b)/c = 1 + (\alpha - 1)b/c$. Hence δ only depends on α and the ratio of the exponents of the transformations φ and ψ.

Summarising Section 3, we have proved that power law transformations φ and ψ yield a Lotkaian IPP with exponent δ as in (33) if the original IPP is Lotkaian with exponent α. As shown in (33), evidently, also the exponents of the power law transformations are involved. Proposition 3.2 shows that δ and α relate as the exponents of the power law transformations in the sense that (in)equalities between δ and α are equivalent with similar (in)equalities between the exponents of the power law transformations.

In the next section we will discuss some (theoretical) applications.

4. Applications

1. No sources are destroyed or created but one has that items can be destroyed (example: no nodes in a network are destroyed or created but one has the destruction of some in-links). Here $A = b = 1$ in (23), clearly. We can assume that the destruction of items follows a random sample in the items, hence sources with a large number of items have a higher probability for an item deletion, the probability being proportional to the source's size. This implies $c = 1$, $0 < B < 1$ in (24) (B being 1- sample probability (for destruction)). In this case we have $\delta = \alpha$ by (33) and (34) is an equality since $b = 1$. We hence refind Lotka's law with the same α.
2. No sources are deleted (destroyed) or created ($A = b = 1$) but in a large source, items are deleted more than proportional to the source's size. Now we have $0 < c < 1$ and B must be choosen small enough to yield less items. Hence (33) yields $\delta > \alpha$. Indeed

$$\delta = \frac{\alpha + c - 1}{c} > \alpha \tag{40}$$

iff

$$\alpha + c - 1 > \alpha c$$

iff

$$(c - 1)(\alpha - 1) < 0$$

which is correct since $\alpha > 1$ and $c < 1$. Now we experience higher exponent values in Lotka's law (34) and again the result is exact since $b = 1$.

3. No sources are destroyed or created ($A = b = 1$) but items are destroyed preferably from low-item sources: $c > 1$ and B must be taken small enough (and certainly <1) so that we have less sources. The same argument as in 2 now yields $\delta < \alpha$.
4. The same result ($\delta < \alpha$) was already found in Egghe (2003, 2004) in case of positive reinforcement (and again no sources are destroyed or created: $A = b = 1$): now $c > 1$ and B large enough to have more items. Again, as above, we have $\delta < \alpha$, as already found in Egghe (2003, 2004). In Egghe and Rousseau (2006a), the same model was used and in addition one supposed $B > 1$ yielding (see Theorem 3.1), besides $\delta < \alpha$, that the minimal j^*-value $a^* = \varphi(a) = Ba^c$ is strictly larger than a, the minimal j-value in the original IPP. Repetition of this transformation leads to IPPs without low productive sources, hence explaining why these IPPs have smaller ($\delta < \alpha$) Lotka exponents: see Egghe and Rousseau (2006a) for examples (cities/villages, countries and database sizes).

An extension of these results is obtained by considering other possible parameter values, but now leading to increased source and item totals.

5. Let us have $0 < c \leq 1$ (as in 1 and 2). Source deletion: let $b > 1$ and $A \in]0,1[$ such that (see (23))

$$\psi(T) = AT^b = T^* < T$$

Then it is obvious from (33) and (40) that $\delta > \alpha$. The same is true for source creation ($b > 1$ and $A > 0$ such that $AT^b = T^* > T$).

6. If $c > 1$ and if we have source destruction such that $b < 1$ (and A such that $\psi(T) = AT^b = T^* < T$) we have, by the argument in 3 and by (33) that $\delta < \alpha$. If $0 < c < 1$ we have no conclusion.

An example is given in Rosen and Resnick (1980) on the distribution of city sizes. Here one has the ambiguity of the definition of "city". One can use urban places, legal cities or urban agglomerations. In Rosen and Resnick (1980) one finds an increase of Zipf's exponent β (there called the Pareto exponent) when going to the larger scale cities such as urban agglomerations. This boils down to a decrease of the Lotka exponent as indicated here ($\delta < \alpha$). This is because of the inverse relation (20) (and similar for δ).

In a way, this result generalizes 3 and 4 as well as Egghe (2003, 2004) and Egghe and Rousseau (2006a) (in the latter article the number of cities remains unchanged).

7. Other examples of $\delta < \alpha$ or $\delta > \alpha$ can be constructed based on given parameter values A, B, b, c and giving further insight in the dynamics (evolution) of IPPs. We note, as in Egghe and Rousseau (2006a), that the given models of source/item creation present a non-stochastic form of general "Success-Breeds-Success" (SBS) principles for this, see Egghe (2005).

Acknowledgements

The author is grateful to Prof. Dr. V. Cothey for mentioning the problem of modelling network evolutions and to profs. Dr. V. Cothey and R. Rousseau for interesting discussions on the topic of this paper. The author is also grateful to two anonymous referees for giving good suggestions to improve this paper, both in content and in style.

References

Egghe, L. (2003). Positive reinforcement and 3-dimensional informetrics. In J. Guohua, R. Rousseau, & W. Yishan (Eds.), *Proceedings of the ninth international conference on scientometrics and informetrics* (pp. 47–54). Beijing (China): Dalian University of Technology Press.

Egghe, L. (2004). Positive reinforcement and 3-dimensional informetrics. *Scientometrics*, *60*(3), 497–509 (*Scientometrics*, *61*(2), 283, 2004 (Correction)).

Egghe, L. (2005). *Power laws in the information production process: Lotkaian informetrics*. Oxford, UK: Elsevier.

Egghe L. (2007). The theorem of Fellman and Jakobsson: A new proof and dual theory (preprint).

Egghe, L. & Rousseau, R. (1990). Introduction to informetrics. Quantitative methods in library, documentation and information science. Amsterdam, The Netherlands: Elsevier.

Egghe, L., & Rousseau, R. (2006a). Systems without low-productive sources. *Information Processing and Management*, *42*(6), 1428–1441.

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

Fellman, J. (1976). The effect of transformations on Lorenz curves. *Econometrica*, *44*(4), 823–824.

Jakobsson, U. (1976). On the measurement of the degree of progression. *Journal of Public Economics*, *5*, 161–168.

Rosen, K., & Resnick, M. (1980). The size distribution of cities: An examination of the Pareto law and primacy. *Journal of Urban Economics*, *8*, 165–186.

Available online at www.sciencedirect.com
ScienceDirect
Journal of Informetrics 1 (2007) 131–144
Journal of INFORMETRICS
An International Journal
www.elsevier.com/locate/joi

Generating overview timelines for major events in an RSS corpus

Rudy Prabowo [a,*], M. Thelwall [a], Mikhail Alexandrov [b]

[a] *School of Computing and Information Technology, University of Wolverhampton, Wulfruna Street, WV11SB Wolverhampton, UK*
[b] *Autonomous University of Barcelona, Barcelona, Spain*

Received 31 August 2006; received in revised form 19 October 2006; accepted 23 October 2006

Abstract

Really simple syndication (RSS) is becoming a ubiquitous technology for notifying users of new content in frequently updated web sites, such as blogs and news portals. This paper describes a feature-based, local clustering approach for generating overview timelines for major events, such as the tsunami tragedy, from a general-purpose corpus of RSS feeds. In order to identify significant events, we automatically (1) selected a set of significant terms for each day; (2) built a set of (term–co-term) pairs and (3) clustered the pairs in an attempt to group contextually related terms. The clusters were assessed by 10 people, finding that the average percentage apparently representing significant events was 68.6%. Using these clusters, we generated overview timelines for three major events: the tsunami tragedy, the US election and bird flu. The results indicate that our approach is effective in identifying predominantly genuine events, but can only produce partial timelines.

Keywords: Feature selection; Clustering; Overview timeline

1. Introduction

The task of identifying significant events from real time news feed data is a standard one in data mining and event detection and tracking (Allan, Papka, & Lavrenko, 1998b; Yang, Pierce, & Carbonell, 1998). The Internet now hosts a range of readily accessible information formats that are new candidates for event detection, and these may come to replace or supplement traditional types, or may give rise to new event detection applications. Really simple syndication (RSS) is one such technology and has already become a widely used standard: it allows blogs and news sources to post-timely information to subscribers, for example, hourly or daily summaries of the most recent updates. RSS feeds have great potential to be used for public-opinion gathering (Glance, Hurst, & Tomokiyo, 2004; Gruhl, Guha, Liben-Nowell, & Tomkins, 2004), mainly because of the large numbers of blog authors maintaining sites with RSS feeds, although bloggers are not typical citizens (Adar, Zhang, Adamic, & Lukose, 2004; Lin & Halavais, 2004) and have a wide variety of motives (Herring, Scheidt, Bonus, & Wright, 2004). In addition, the concise RSS formats allow relatively low-bandwidth data gathering, even for a large number of different sources. When a major (or world) event, such as the Asian tsunami (26/12/2004), occurs, RSS feeds could therefore be used to generate an overview timeline of the event.

Our contributions are to develop an automatic method to achieve the following using RSS data.

* Corresponding author. Tel.: +44 1902 518584; fax: +44 1902 321478.
E-mail addresses: Rudy.Prabowo@wlv.ac.uk (R. Prabowo), Mike.Thelwall@wlv.ac.uk (M. Thelwall), dyner1950@mail.ru (M. Alexandrov).

1751-1577/$ – see front matter
doi:10.1016/j.joi.2006.10.002

(1) Find daily sets of significant terms (either nouns or noun phrases) which maybe associated with important events, i.e. the most discussed happenings (Section 3).
(2) Use the significant terms to build a set of (term–co-term) pairs and cluster the pairs. The clusters are our candidates for the day's significant events (Section 4).
(3) Generate overview timelines for major events by sorting the clusters by date (Section 5).

In this paper, we are primarily interested in the precision of the clusters in (2). More specifically, we assess the extent to which human judges agree that the automatically generated clusters genuinely describe a single event. A human-based evaluation was important to discover whether the results could be understood by potential end users, i.e. human interpreters.

To illustrate the timeline generation, three major events, 'tsunami tragedy', 'US election' and 'bird flu spreading', were selected as our case studies. Tables 5–7 show the generated timelines for each major event. Each timeline refers to one particular major event along with many related, subsequent events.

2. Related work

This section reviews existing work in the area of (1) term selection, (2) topic and event detection and tracking (TDT) and (3) timeline generation.

2.1. Term selection

Given a set of terms (e.g. words, word stems, nouns and noun phrases) in a document collection, selecting the most significant terms is the first step. This stage is common to a range of specific tasks, including information retrieval (IR), automatic text classification and time series analysis. The selected terms represent document features in the form of a term vector: a list of the most significant terms from the document and their frequencies in the document.

Either *tf·idf* (Salton & McGill, 1986) or *lnu* weighting (Singhal, Buckley, & Mitra, 1996) can be applied to assign each term a value which estimates its significance. These formulae take into account both local and global term frequency. In IR, the assigned value is then used as a starting point to: (1) compute the similarity between the documents available in the corpus and a user query and (2) rank search results in order of relevance to a user query. In an ideal scenario, each term should be assigned a degree of significance, such that an IR system can achieve a high precision level at 100% recall (Baeza-Yates & Ribeiro-Netto, 1999; Belew, 2000). It is not suitable for our event detection task, however, because important events may be identified through single highly significant terms (Swan & Allan, 2000).

The three formulae that have previously been used to identify significant events in blog or RSS corpora are variants of *tf* and *tf·idf* (Gruhl et al., 2004; Glance et al., 2004; Thelwall, Prabowo, & Fairclough, 2006). The formulae do not, however, depend on the full document space, but on a fixed time period as a time window of observations, and are used to quantify the 'burstiness' of a term within the fixed, short time period, i.e. the degree of importance of terms within the time period. The result changes if another time window is used, for example, 1 week earlier or later. While this feature is useful to keep track of the burstiness of terms for different time windows, it is less suitable for the initial identification of significant events. For this, the degree of significance of a term over a long period of time is required, e.g. 1 year.

The commonly used formulae for identifying significant terms in the area of automatic text classification are: χ^2, Mutual Information (MI) and Information Gain (I) (Sebastiani, 2002). Swan and Allan (2000) use a χ^2-based method to determine the degree of significance of terms on given dates. Nevertheless, it is not yet clear whether this is the best method for all types of data, and in the context of RSS, it is also worth harnessing the power of the Information Gain method (Prabowo & Thelwall, 2006).

2.2. TDT

In the context of the TDT task, an event is defined to be something that happens at a specific time and place, whereas a topic is defined more widely as a seminal event or activity, along with all directly related events and activities (Allan, Carbonell, Doddington, Yamron, Yang, 1998a). The term 'story'; is often used to describe the natural unit of text in which the information arrives, such as a single newswire report. The topic detection and tracking tasks focus on

the identification of topics across stories. The event detection task is to cluster together stories that refer to the same event. The event-tracking task is concerned with assigning each incoming new story to the most appropriate event that it discusses, if any. Otherwise, the story is assumed to detect a new event (Allan et al., 1998b). In contrast, our task focuses on generating an overview timeline with respect to one particular major event and its related, subsequent events.

To detect and track events, the following method is used. Given a set of stories, assign each term found within each story a weight. The weighted terms of a story are stored as a vector, and regarded as the representation of the story. In the case of event detection, given a set of terms, stories that discuss the same event are clustered together. In the case of event tracking (1) the similarity between the term vector of a new story and all the term vectors of the existing stories is computed and (2) the most appropriate existing event is assigned to the new story, if any. Otherwise, a new event is recorded.

2.3. Timeline generation

As timestamped data, such as blog data and news articles, have become available, the timeline generation task has attracted a number of researchers. Glance et al. (2004) used blog data to carry out topic mining, detect key persons and produce a timeline for a topic or a key person. The method used is as follows: (1) select a set of significant phrases; (2) cluster all the selected phrases. Two phrases are clustered together, if the cosine similarity of their occurrence is greater than a threshold. Each cluster represents a topic; (3) generate timelines for the topics.

Swan and Allan (2000) focused on assigning each significant term a time period by using χ^2-tests. Swan and Allan (2000) ranked the significant terms according to their χ^2 values in a descending order, i.e. the term with the largest χ^2 value was at the top of the list. Then, they compared the time period of a term with all the lower ranked features. If their time periods overlapped, they carried out χ^2-tests to determine whether the two features were dependent, If yes, they marked the features as potential members of a cluster. Finally, they carried out a hierarchical agglomerative clustering on the marked features. The clusters were evaluated against a set of predefined topics (discussed in detail in Section 4.5).

Smith (2002) extracted a number of place names from historical documents and assigned each place name a date. Then, Smith (2002) ranked the collocations of place name and date pairs according to their log-likelihood values. The significant collocations can be displayed in timelines.

In contrast, we do not focus on a topic, but on the generation of a timeline for the wider concept of a major event. The generated timeline contains not only the major event, but also its related events. Our work is more related to Smith (2002), than to Glance et al. (2004) and Swan and Allan (2000). Our approach, however, is more general. We did not only exploit place names, but also other phrases as a way to build a set of (term–co-term) pairs and to cluster the pairs. The TDT corpus was built for single topic/event detection exercise, but was not annotated for the purpose of evaluating overview timelines of an event and its related events. We therefore could not use the TDT corpus for our experiment. Instead, we use RSS data, which is the type of data that we believe can be usefully exploited for timeline generation.

3. Significant term selection

The selection of significant terms was conducted in three stages.

(1) RSS items were collected and the text found within the items was processed;
(2) χ^2 and Information Gain (I) values were computed;
(3) A set of significant terms was selected.

3.1. Pre-processing texts

The following procedure was used to pre-process the texts found within a set of RSS items.

(1) Mozhdeh (Thelwall et al., 2006) was used for collecting data. The system monitored 19,587 RSS feeds hourly (daily for infrequently updated feeds). Each feed returns a set of items, with each item containing a separate set of information. The system stored each new item found.

Table 1
A 2 × 2 contingency table

	term_i	$\overline{\text{term}_i}$
date_j	a	b
$\overline{\text{date}_j}$	c	d

(2) For each item, the title, description and publication date were automatically extracted and stored in a plain text file. Each publication date was converted into its associated GMT time by using a date converter. For each file, a part of speech (POS) tagger (Brill, 1992) and noun phrase (NP) chunker (Ramshaw & Marcus, 1995) were used to tag and chunk the item texts, extracting nouns and noun phrases, including proper nouns. The tagger achieved 95% precision (Brill, 1992) and the chunker 93% (Ramshaw & Marcus, 1995). The tagger and chunker, which run on an intel P4 3.2 GHz, can process about 650K items per day. They could therefore operate in real time.
(3) Three inverted files were built.
   (a) TermItem: Used to determine to which item each term belongs;
   (b) ItemRSSFeed: Used to determine to which RSS feed each item be longs;
   (c) ItemDate: Used to determine when each item was posted in GMT time (i.e., publication date).

3.2. Computing χ^2 and Information Gain (I)

Given a term, term_i and a publication date, date_j, the 2 × 2 contingency table used for calculating a χ^2 value is constructed as follows (Table 1).

The χ^2 value of term_i with regard to date_j was calculated as follows:

$$\chi^2 = \sum_k \frac{(\mathrm{O}_k - \mathrm{E}_k)^2}{\mathrm{E}_k} \tag{1}$$

- $k = \{a, b, c, d\}$;
- O_k is the observed frequencies in a 2 × 2 contingency table with respect to term_i and date_j;
- E_k is the expected frequencies of O_k with respect to term_i and date_j.

A Yates continuity correction was applied to each χ^2 calculation, as the degree of freedom is 1. Large χ^2 values suggest that term_i and date_j are dependent upon each other.

Let D be a binary date variable, $\{d, \bar{d}\}$ representing the presence or absence of a date and T be a binary term variable, $\{t, \bar{t}\}$ representing the presence or absence of a term. Information Gain, $I(D; T)$ was computed based on $H(D)$, an entropy value for D and $H(D|T)$, the conditional entropy of D given T and is defined as follows:

$$\begin{aligned} I(D;T) &= H(D) - H(D|T) \\ I(D;T) &= \left\{ -\sum_{j=d,\bar{d}} p(j)\cdot\log_2 p(j) \right\} - \left\{ p(t)\cdot\sum_{j=d,\bar{d}} H(j|t) + p(\bar{t})\cdot\sum_{j=d,\bar{d}} H(j|\bar{t}) \right\} \\ I(D;T) &= -\left\{ p(d)\cdot\log_2 p(d) + p(\bar{d})\cdot\log_2 p(\bar{d}) \right\} - \left\{ p(t)\cdot\sum_{j=d,\bar{d}} H(j|t) + p(\bar{t})\cdot\sum_{j=d,\bar{d}} H(j|\bar{t}) \right\} \end{aligned} \tag{2}$$

The conditional entropy of D given T was computed as follows:

$$\sum_{j=d,\bar{d}} H(j|t) = -\left\{ p(d|t)\cdot\log_2 p(d|t) + p(\bar{d}|t)\cdot\log_2 p(\bar{d}|t) \right\} \tag{3}$$

$$\sum_{j=d,\bar{d}} H(j|\bar{t}) = -\left\{ p(d|\bar{t})\cdot\log_2 p(d|\bar{t}) + p(\bar{d}|\bar{t})\cdot\log_2 p(\bar{d}|\bar{t}) \right\} \tag{4}$$

In information theory (Shannon & Weaver, 1963), $I(D; T)$ is used to measure the average reduction in uncertainty about D that results from learning the value of T (MacKay, 2003). In our context, we (1) apply this notion of average reduction in uncertainty to determine the degree of closeness between D and T by learning the presence and absence of T in each RSS item, with regard to the $H(D)$ of one particular publication date and (2) use the value of $H(D|T)$ to quantify the degree of uncertainty that a term T is significant on a date D. The smaller a $H(D|T)$ value is, the higher the degree of certainty that T is a good indicator for D. For $H(D|T) = 0$, the degree of uncertainty in learning the value of D is 0, which means that T is highly indicative for a date D.

The time complexity of χ^2, as well as Information Gain (I) is $\mathcal{O}(\mathcal{T} \cdot \mathcal{D})$, where $\mathcal{T}$ is the number of terms found within a set of RSS items and $\mathcal{D}$ is the number of publication dates used.

3.3. Selecting significant terms

From the 19,857 monitored RSS feeds, 880,536 different items were extracted between 01/01/2004 and 28/02/2005. A total of 1,736,715 unique terms were extracted from a total of 413 publication dates (some items were clearly old when collected). The 127,859 were single word terms and 1,608,856 were multi word terms. From these data, 2,912,581 term–date pairs were generated. Each term–date pair was assigned χ^2 and I values as described above. Our previous evaluation results (not using human evaluators) suggest that χ^2 would be the best of the three methods (Prabowo & Thelwall, 2006). Nevertheless, it is far from perfect as extremely high values can still occasionally be assigned to relatively insignificant terms. The results also showed that χ^2 and I had a strong degree of agreement when judging the term significance and I can aggressively remove some insignificant terms. In attempt to extract a high proportion of genuinely significant terms, only those which were judged to be significant by both χ^2 and I were selected, a total of 684,431.

4. Term clustering

This section discusses the way we automatically clustered together related terms and manually evaluated the clusters.

4.1. Clustering procedure

The 'features' of a document, i.e. its most significant terms, ideally should represent the essence of the document. Given a collection of documents, document clustering is the operation of grouping together similar (or related) documents, with respect to their features (Baeza-Yates & Ribeiro-Netto, 1999).

In our case, the clustering of items can be computationally intractable because of the size of the matrix: 880,536 items·684,431 features. Even with daily clustering, there are still thousands of items to process per day. To overcome the computational issue, a matrix of associations between significant terms only can be used as an alternative. This approach, however, does not take the context into account. For these reasons, we exploited term co-occurrence to provide context to a set of significant terms. The clustering procedure is described below.

For each daily set of significant terms, we extracted the terms which co-occurred with each significant term in the same sentence and generated (term, co-term) pairs. The co-terms do not necessarily need to be significant terms. Thus, a significant term was the trigger to initiate a context, which was represented by (term, co-term) pairs. To ensure that the clustering algorithm would not modify the context, unless other significant terms were linked to the existing co-terms, we used each significant term as an attribute and treated the co-terms as objects, i.e. we clustered the co-terms, rather than the significant terms. No clustering algorithm would have been necessary, if only one event occurred on a single day. This, however, was typically not the case.

Given a set of (term, co-term) pairs from a single day, only the pairs that occurred at least twice were selected (pair count threshold = 2) to remove infrequent pairs. Next, a covariance matrix, in form of a higher triangular (HT) form, was created. The cosine measure was used for measuring the closeness between two objects (co-terms). Finally, the MajorClust algorithm (Stein & Niggemann, 1999) was used to cluster the pairs.

Let G be a graph; $C = \{C_1, C_2, \ldots, C_n\}$ be the decomposition of G; $|C_i|$ be the number of nodes in C_i; λ be the minimum number of edges that must be removed to make G an unconnected graph. To get the best decomposition of

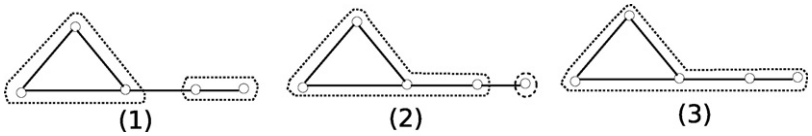


Fig. 1. An example to illustrate $\Lambda(\mathcal{C})$.

the graph, G, we need to compute and maximise the value of the weighted partial connectivity, $\Lambda(\mathcal{C})$:

$$\Lambda(\mathcal{C}) = \sum_{i=1}^{n} |C_i| \cdot \lambda_i \tag{5}$$

Let us use Fig. 1 as an example to illustrate $\Lambda(\mathcal{C})$.

The graph depicted above can be decomposed into three different possibilities. For (1), $\Lambda = 3{\cdot}2 + 2{\cdot}1 = 8$. For (2), $\Lambda = 4{\cdot}1 + 1{\cdot}0 = 4$. For (3), $\Lambda = 5{\cdot}1 = 5$. (1) yields the largest Λ value. Therefore, (1) is regarded as the best structural division of subgraphs. A $\Lambda(\mathcal{C})$-based clustering algorithm avoids long chaining effects, as illustrated in (3). The algorithm can be adjusted to handle a weighted graph.

In our experiment, we used the MajorClust algorithm, which follows the notion of $\Lambda(\mathcal{C})$ optimisation, to cluster objects with respect to their attributes. At the initial step, the algorithm assigns each object to its own cluster. In the next iterations, each object is assigned to a cluster to which the closeness of the object is stronger than to other clusters. If more than one cluster exists, one of them is randomly selected (Stein & Niggemann, 1999; Alexandrov, Gelbukh1, & Rosso, 2005). This iterative optimisation strategy leads to the local optimum of the best possible division of objects. The algorithm does not need to know the number of clusters (k) in advance and can efficiently process a large graph, as it only looks for a local optimum and does not carry out a hierarchical clustering. These characteristics are suitable for our task for the following reasons. A collection of terms must be efficiently processed. A hierarchical, agglomerative clustering may offer a global optimum, but would be computationally expensive and we would also need to estimate the level where the dendogram is to be cut to get the desired clusters. A localised variant of the k-means algorithm may run faster than MajorClust, but the MajorClust runs reasonably fast, does not suffer from chaining effect and suits our data particularly well due to its graph-based approach. The time complexity of the algorithm is $\mathcal{O}(E{\cdot}C_{\max})$, where E is the number of edges representing the connections between clusters and objects and $C_{\max}$ is the largest cluster in the graph.

The outputs were a set of clusters, each of which contained at least one co-term. We also mapped each cluster into its associated significant terms. By doing this, we can keep the context in which a co-term was used. For example, cluster 7 (Table 6), contains the co-term, 'Japan'. The two significant terms 'bird flu fear' and 'bird flu virus' determine the context in which the co-term 'Japan' occurs.

4.2. Evaluation procedure

The following procedure was used to evaluate the automatically generated clusters described above. Recall that we carried out a human-based evaluation, because we wanted to produce information that was meaningful to human interpreters and hence wanted to know whether the clustering procedure was producing information that could be understood by potential end users.

(1) A topic is defined as a seminal event, along with all directly related events (NIST Speech Group, 2005). We adopt the above definition and interpret it in a holistic manner, i.e. a topic is an abstraction of a series of events which are represented by groups of significant terms, which means that a whole has more value than its parts. In this respect, any significant term/an event should not be regarded as the full representation of a topic, but only as a single part of a whole.

From the 24,353 clusters, we manually selected clusters which were associated with 20 large topics (e.g. bird flu, tsunami, US presidential election 2204, global warming). As explained above, each topic is the abstraction of a series of events. A topic was chosen if there were at least 50 clusters which described an important/major event and its subsequent, related events. We used 20 large topics as a starting point to select a set of clusters which signified day's events. For each group, we then randomly selected 5 clusters; the 100 'real clusters'.

(2) From the 100 real clusters, we generated a list of terms which contained all the terms in the clusters. Each term was assigned an index. From this term list, we randomly generated 100 pseudo clusters, each of which had the same length as a counterpart in the real set. These 100 'fake clusters' formed the control group for the experiment.
(3) We shuffled the 100 real and 100 fake clusters and put them into a list. We then asked 10 staff members in our school to judge whether the clusters were good or bad and the results were returned anonymously to the first author. In cases where they were not sure we asked them to abstain. Here, a good cluster is defined to be a cluster which contains contextually related terms, i.e. the terms seem to relate to the same event.

4.3. Evaluation results

We used χ^2-tests to determine whether each individual assessor could tell the difference between the 100 real and the 100 fake clusters. The 10 χ^2 values range from 30.16 to 107.89, all greater than the critical value $\chi^2_{0.01} = 9.21$. Hence, there is a strong evidence to reject H_0. This means that for all the 10 assessors the proportion of good, bad and abstention for the 100 real clusters was significantly different from those for the 100 fake clusters. For the 100 real clusters, the average proportions of good, bad and abstention are 68.60, 14.40 and 17 and for the 100 fake clusters, 16.70, 64.10 and 19.20 and so the effectiveness of the clustering method is clear, although it is not perfect.

In addition, Friedman tests were carried out to determine whether the proportion of good clusters was significantly higher for the real than for the fake clusters across the 10 assessors as a group. Both the good and bad proportions of the real and fake clusters yield the same result: $\chi^2 = 10$ ($p = 0.002$). This confirms that the assessors could differentiate the real from the fake clusters a significant part of the time.

We encountered problems in the evaluation: to find a group of people who had broad knowledge about news stories. Two assessors recorded a very high level of abstentions (33 and 47). This may be because the assessors were not familiar with the particular news stories. Thus, the evaluation results may well underestimate the real performance of our clustering approach, from the perspective of an expert interpreter. Nevertheless, the results clearly indicate that the 100 real clusters are significantly better than the 100 fake clusters.

4.4. Further analysis of results

To better measure the performance of our clustering approach, we counted the number of good clusters for each topic, with regard to the 10 assessors and averaged the 10 numbers, as formally defined below.

$$G_{\text{avg}}(\tau) = \frac{\sum_{a=1}^{n} G(\tau, a)}{n} \tag{6}$$

Here, a is an assessor, $n = 10$ is the total number of assessors and τ is a topic. The value $G(\tau, a)$ is the number of good clusters with regard to a topic, τ and an assessor, a. The $G(\tau, a)$ values range from 1 to 5 and $G_{\text{avg}}(\tau)$ is the average of all the $G(\tau, a)$ values. Each topic is represented by five clusters, as stated in Section 4.2. The judgement of an assessor was not taken into account if they abstained from all five clusters, which we took to mean that they were not familiar with the topic or were reluctant to do the evaluation. Hence, this method avoids the hypothesised bias from some of the assessors. Given a topic, the best mark would be 5, meaning that an assessor judged all the five clusters related to the topic to be genuine. The lowest mark would be 1, meaning that an assessor only judged one cluster to be genuine. Table 2 shows the $G_{\text{avg}}(\tau)$ of all 20 topics.

The average of all the $G_{\text{avg}}(\tau)$, listed in Table 2 was 3.57 (71.4%). Analogously, we adapted and applied Eq. (6), to compute $B_{\text{avg}}(\tau)$ for bad clusters and $A_{\text{avg}}(\tau)$ for neither good nor bad. The average of all the $B_{\text{avg}}(\tau)$ was 0.75 and the

Table 2
The $G_{\text{avg}}(\tau)$ of all the 20 topics

τ	$G_{\text{avg}}(\tau)$	τ	$G_{\text{avg}}(\tau)$	τ	$G_{\text{avg}}(\tau)$	τ	$G_{\text{avg}}(\tau)$	τ	$G_{\text{avg}}(\tau)$
τ_1	4.22	τ_5	4.10	τ_9	4.30	τ_{13}	4.30	τ_{17}	3.25
τ_2	3.40	τ_6	3.90	τ_{10}	2.56	τ_{14}	3.00	τ_{18}	3.13
τ_3	3.30	τ_7	3.11	τ_{11}	3.33	τ_{15}	3.80	τ_{19}	3.78
τ_4	4.40	τ_8	3.80	τ_{12}	3.70	τ_{16}	2.78	τ_{20}	3.20

Table 3
The number of real clusters judged bad by the assessors for different level of agreement

#Assessors (level of agreement)	#Real clusters judged bad	Cluster-id
5	6	11, 27, 37, 69, 87, 161
6	3	11, 27, 161
7	2	11, 27
8	1	11
9	0	–
10	0	–

average of all the $A_{\text{avg}}(\tau)$ is 0.68. Taking all the assessors's judgements into account, then the following results were obtained. The average of all the $G_{\text{avg}}(\tau)$ was 3.43 (68.6%). The average of all the $B_{\text{avg}}(\tau)$ was 0.72 and the average of all the $A_{\text{avg}}(\tau)$ was 0.85. Clearly, the average of all the $G_{\text{avg}}(\tau) = 3.43$ indicates a slight bias (2.8%) from some assessors.

In addition, we counted the number of real clusters which were judged bad by the assessors at the x level of majority, where $x = \{5, \ldots, 10\}$. Here, 'x' means that there are at least x assessors who think that a real cluster is a bad cluster. Table 3 lists the results. The six real, bad clusters were individually analysed to find the reason why they were judged bad. Four assessors who judged the six real clusters as bad clusters were willing to give their reasons for their judgement. Table 4 lists the six real clusters. The summary of all the RSS items which are associated with the six real clusters is listed below.

- 11: The experiment which shows that children are more vulnerable than adults, in regard to the adverse effects of air pollution.
- 27: The Vietnamese government took action to prevent bird flu from spreading in Yunnan province in China.
- 37: An experimental finding that shows that ethanol can promote cancer progression.
- 69: Vietnam deployed riot police at bird flu check points around Ho Chi Minh city.
- 87: The basic right to choose with whom someone wants to create their children.
- 161: The shifts in the apparel trade roil [i.e. agitate] the global economy which can threaten the living standards of poor nations.

The reasons for judging the clusters as bad clusters are listed below.

- Cluster 11 was judged bad, because the assessors thought that the term, 'adults', is irrelevant to the adverse effects of air pollution. In this case, only two term co-term pairs were used, as the associated RSS items are quite sparse. Thus, it leads the clustering algorithm to cluster the related co-terms together.
- Clusters: [27, 37, 161] were judged bad, because the assessors thought that the cluster members are contextually not related. This is due to the input data for the clustering algorithm, which was too sparse, even though the associated RSS items contain sufficient data. We only used the co-terms which can be found in the same sentence (a fixed window of observation). We might be able to avoid this problem by widening the window of observation, but this might not work well in another case, as some clusters may contain too many unrelated terms.
- Cluster 69 was judged bad, because the assessors thought that the term, 'riot police', is irrelevant to the other cluster members. There should be no connection between 'bird flu' and 'riot police'. This problem was anticipated from

Table 4
The six real clusters judged bad

Cluster-id	Cluster members
11	Adults, adverse-effects, air pollution
27	Bird flu intrusion, Yunnan-guards
37	Cancer progression, ethanol
69	Bird flu checkpoints, Hanoi Vietnam, Ho Chi Minh city officials, riot police
87	Basic human rights, children, strip
161	Apparel trade roil global economy, shifts, unravels

the beginning; it is difficult to judge two contextually related terms, unless the assessors know the event with which the cluster is associated.

- Cluster 87 was judged bad, because the assessors thought that the term, 'strip', is irrelevant to the other cluster members. This is due to a parsing error, as the POS tagger tagged the term as a noun.

In summary, given the six real clusters, there is one real cluster (cluster 69) which is actually a good cluster, in the sense of containing sufficient information to indicate a single event, but was judged bad by the assessors. The root cause of cluster 69 being judged bad is probably that the event described was relatively minor and quite specific with an unusual collection of terms. It is likely that the assessors did not know of the event or had forgotten it. The other five clusters are bad clusters, due to data sparseness, the length of window of observation and parsing errors.

We also found a chain effect of clustering, i.e. two unrelated events clustered together because the input data misleads the clustering algorithm. The following two examples illustrate the problem. Cluster 5 (in Table 5) grouped two unrelated events together. The term, 'clean water' in cluster 5 was significant on 28/12/2005 and became the trigger which initiated two different contexts. One was in the context of tsunami tragedy and another one was in the context of John Kerry's speech about clean water. The term, 'bloomberg' (a media service) in cluster 1 (in Table 6) became the trigger which initiated one context, i.e. world crisis, which subsumed the two terms, 'fuel costs' and 'bird flu'.

4.5. Comparison with existing work

Swan and Allan (2000) focussed on generating the overview timelines of significant terms and carrying out a clustering on the significant terms. The Kappa statistic κ (Cohen, 1960; Siegel & Castellan, 1988) was used to measure pairwise agreement among assessors. For $0.67 \leq \kappa < 0.8$, a tentative conclusion can be drawn. For $\kappa \geq 0.8$, a definite

Table 5
An excerpt of an overview timeline for the 'tsunami' event

Id	Date	Cluster
1	2004-12-26	Asian-nations years-triggers-tsunami earthquake
2	2004-12-27	Asian-death-toll Asian-disaster death-toll earthquake-tsunami
3	2004-12-27	Australian-red-cross Asia-quake tsunamis-appeal
4	2004-12-28	Desperate-refugees tsunami death-toll-climbs
5	2004-12-28	Asian-tsunami disasters senator-Kerry spread aftermath clean-water
6	2004-12-29	Asian-earthquake tsunami pledge UK-government victims
7	2004-12-29	Basic-equipment Indonesian-tsunami monitoring-system
8	2004-12-30	Bush-administration pledges support tsunami-aid
9	2004-12-30	Aid-efforts Asian-tsunami pledge victims
10	2004-12-30	Tsunami-scientists earthquake-prone-nation public-safety-officials
11	2004-12-30	Banda-Aceh-Indonesia tsunami death-toll-jumps
12	2004-12-31	Asian-tsunami-tragedy aircraft-carrier-battle-groups navy-battle-groups tsunami-relief
13	2004-12-31	Banda-Aceh-Indonesia relief-efforts stricken-area victims aid tsunami-toll-climbs
14	2004-12-31	Basic-requirements basic-services devastating-tsunamis tragedy southeast-Asia
15	2005-01-01	Pope-John-Paul-II special-mass-early-Saturday tsunami-victims
16	2005-01-01	Diseases biggest-threat tsunami-survivors
17	2005-01-07	Aceh-destruction tsunami-destruction Annan
18	2005-01-10	Aid-money tsunami-relief public-tracking-system
19	2005-01-10	Aceh board-crashes relief-operation helicopter-crash
20	2005-01-22	Aceh-province early-warning-system tsunami lives

Table 6
An excerpt of an overview timeline for the 'bird flu' event

Id	Date	Cluster
1	2004-12-02	Bloomberg fuel-costs bird-flu
2	2004-12-08	Pandemic Asian-bird-flu virus
3	2004-12-09	Bird-flu-pandemic who governments
4	2004-12-10	Youngest-sars rapid-bird-flu-test Beijing Hong-Kong Xinhuanet-scientists
5	2004-12-18	Bird-flu virus-antibody blood-samples
6	2004-12-18	Japan outbreak bird-flu-virus culling
7	2004-12-19	Bird-flu-fear bird-flu-virus Japan
8	2004-12-20	Human-flu-virus bird-flu pandemic
9	2004-12-22	Avian-influenza first-human-infection bird-flu
10	2004-12-23	Dangers disease human large-scale outbreaks bird-flu
11	2004-12-31	Deadly-bird-flu-virus poultry fresh-outbreaks
12	2005-01-07	Bird-flu-case Vietnam girl
13	2005-01-08	Bird-flu-intrusion Yunnan-guards bird-flu-spreads
14	2005-01-21	Human-bird-flu-transmission birds virus Vietnam
15	2005-01-23	Bird-flu-deaths human-toll Vietnam possible-global-flu-pandemic
16	2005-01-24	Bird-flu-case bird-flu-infections positive-cases Vietnam-reports
17	2005-01-26	World-bird-flu-fear pandemic deaths
18	2005-01-28	Vietnamese-girls bird-flu southern Vietnam
19	2005-01-29	Bird-flu-checkpoints Hanoi Vietnam riot-police ho-chi-mink-city-officials
20	2005-01-30	Bird-flu-evolution Hanoi WHO-experts

conclusion can be drawn (Eugenio & Glass, 2004). Swan and Allan (2000) provided a list of TDT-2 topics, to four assessors and asked the assessors to assign each generated cluster to one/more of the TDT-2 topics. The assessors were allowed to define their own topic, if they were not happy with the topics provided. The percentage of agreement on how many topics each cluster should be assigned was 73.6%, with $\kappa = [0.045 - 0.315]$, and a $\kappa_{\text{average}} = 0.223$. Swan and Allan (2000) state that the low κ values were due to the notion of topic upon which the four assessors could not agree with each other. This is similar to our case, in the sense that the concept of topic is important, but yet inherently ill-defined. When the assessors were given a set of clusters which were already assigned a pre-defined TDT topic and were only asked to indicate whether they agreed, the percentage of agreement on the assigned topics was 86.7%, with $\kappa = [0.6 - 0.785]$, and $\kappa_{\text{average}} = 0.699$ (Swan & Allan, 2000).

In contrast, we focussed on generating overview timelines for major events from their related, subsequent events. We clustered the co-terms of significant terms, as we wanted to cluster all the terms which are contextually related and used the cluster to signify day's events. We did not ask the assessors to assign a specific topic to each cluster. We carried out the κ test to measure the pairwise agreement among the assessors on judging the 200 clusters. The κ values $= [0.21 - 0.60]$, and $\kappa_{\text{average}} = 0.36$. As all the κ values obtained were <0.67, there is no need to carry out κ test (Eugenio & Glass, 2004), for comparison. Clearly there is a low rate of agreement among assessors.

In our experimental setting, we did not ask the assessors to deal with the notion of a topic. Instead, we asked them to judge whether the cluster contains contextually related terms which may signify a day's significant event. Clearly, our evaluation focuses on the context between cluster elements and not the mappings between a cluster and topics. We deliberately did not give the assessors the RSS items associated with a cluster or a set of pre-defined events, as it would introduce a bias in judging the cluster. This explains the reason for the low κ values obtained, because event knowledge is an additional complicating factor.

5. Clustering results

We use a qualitative approach to investigate how clusters relate to major events. The objective is to gain insights into the types of information indicated by the clusters and how this may vary by major event type. We automatically selected a portion of clusters which were related to three important events, 'tsunami', 'bird flu' and 'US presidential election', which happened in 2004. Each major event was termed 'an initial event', as it was the beginning of a series of 'subsequent events'. In this context, a subsequent event is an event which carries both of the following.

- The essence of an initial event, as it is directly or indirectly triggered by the initial event.
- Its own meaning, which extends the scope of an initial event.

For each initial event, all the clusters which contained the significant term which signified the event were automatically selected: 'tsunami', 'bird flu' and 'election'. For the 'US presidential election', we also selected the clusters which contained the terms, 'George W. Bush' and 'John Kerry', as they were the most influential presidential candidates in the 2004 US election. To illustrate the way in which we analysed each initial event along with its subsequent events, only 20 clusters for each initial event are listed in Tables 5–7.

5.1. Qualitative analysis

The 20 clusters listed in Table 5 shows an excerpt of an overview timeline for the tsunami of 26/12/2005. By manually analysing the RSS items – on each day – in which each term occurred, we found the following points.

Table 7
An excerpt of an overview timeline for the 'US presidential election' event

Id	Date	Cluster
1	2004-01-23	Election Internets
2	2004-01-24	John-Kerry Howard-Dean Iraq
3	2004-01-24	Election President-Bush White-House election national-security
4	2004-03-03	John-Kerry George-Bush Presidential-election
5	2004-05-07	George-Bush Iraq-prison-abuse-scandal re-election-campaign
6	2004-05-11	Impact president-falling-poll-numbers re-election-chances
7	2004-05-19	Bush-White-House credit twist strategically important-states machinery re-election
8	2004-07-06	Presidential-election chances winning
9	2004-07-11	Presidential-election postponement terror-attack elections
10	2004-07-14	Effect case elections plan later-date terror threat
11	2004-10-20	Impact finances presidential-election
12	2004-10-21	Citizens presidential-election country endorsements huge-deal
13	2004-10-28	Presidential-election voting redskins winner
14	2004-10-30	Minds quiet-issue liberals conservatives presidential-election
15	2004-11-01	Presidential-elections outcome year
16	2004-11-03	Outcome George-Bush re-election presidential-election
17	2004-11-06	Electronic-touch-screen voting machines
18	2004-11-15	Prospect election-result president
19	2004-12-14	Americans election real-determining-factor
20	2005-01-20	George-Bush inauguration US-president

- The estimation of the scale of tsunami. Cluster 1 highlighted the estimation of the scale of the initial event that affects five Asian nations.
- Assessing the need to install a tsunami monitoring system. Clusters: [7, 10, 20] were a specific issue about the need of having a tsunami monitoring system to prevent the tragedy in the future.
- The consequence and problems which occurred due to the tsunami. Clusters: [2, 4, 5, 11, 13, 14, 16] described major problems in the aftermath of the tsunami, such as the death toll, disease, refugees and the lack of basic requirements and services, such as clean water. Cluster-id 19 was an isolated incident about a helicopter crash near the Banda Aceh (Indonesia) airport during the relief operations.
- The actions taken to help tsunami victims. Clusters: [3, 6, 8, 9, 12, 18] referred to the Australian red cross appeals for donations and the UK and US government pledges to help the tsunami victims. For example, the US government sent navy aircraft carrier battle groups to deliver relief aid (cluster 12) and a public tracking system was set up to organise aid money (cluster 18). Clusters: [15, 17] highlighted two prominent public figures, Pope John Paul II and Mr. Kofi Annan (the secretary general of the United Nations), who expressed their condolences to the tsunami victims.

The 20 clusters listed in Table 6 shows an excerpt of an overview timeline for the spread of bird flu from 12/2004 to 01/2005.

- The global pandemic hypothesis. Clusters: [2, 5, 8, 9, 10, 11, 14, 15, 17, 20] indicated a predictive assessment from the World Health Organization (WHO) and scientists about the possibility that bird flu could be transmitted from human to human and become a global pandemic.
- The major problems which were raised due to the virus spreading. Clusters: [1, 5, 7, 9, 10, 12, 14, 15, 16, 18] showed the continuing fears and problems faced by the affected Asian countries, i.e. the danger of the bird flu virus for human life and national economies.
- The actions taken against the bird flu virus. Clusters: [3, 11] referred to WHO warnings of a possible bird flu pandemic. Cluster 4 was about a Chinese scientist who carried out a rapid bird flu test. Clusters: [6, 13, 19] referred to the actions which were taken to prevent the virus from spreading.

The two initial events, ‘tsunami’ and ‘bird flu’, are different in the way in which the initial event occurred. Within 01/12/2004–31/01/2005, the tsunami only occurred once, but had a devastating impact. In contrast, bird flu occurred more than once and periodically claimed human lives. Five (clusters: [5, 9, 10, 14, 15]) of the 20 clusters described both the danger of the bird flu virus and the global pandemic hypothesis, as the WHO and scientists on several occasions made predictive assessments, when bird flu claimed human lives. In contrast, in the tsunami example, the clusters were more about the problems raised due to the tsunami attack and the actions in bringing relief aid to stricken areas, organising financial support for reconstruction and addressing critical issues, such as the need for clean water.

Table 7 shows an excerpt of an overview timeline for the ‘US presidential election’ event. The majority of the clusters contained a number of events in which the presidential candidates made a political decision or a political judgement which led to a number of questions – concerning the election – as to whether:

- The influence of Internet on the election had been over hyped (cluster 1).
- John Kerry’s judgement over Iraq was right (cluster 2).
- John Kerry could win (cluster 4).
- Iraq prison abuse could affect the election result (cluster 5).
- President Bush’s falling poll ratings could have an impact on his re election (cluster 6).
- John Kerry’s choice of John Edwards could help or hurt his chances of winning (cluster 8).
- The presidential election should be postponed due to terrorist threats (clusters: [9, 10]).
- The presidential election would have an impact on individual finances (cluster 11).
- The presidential election would have an impact on the citizens of other nations (cluster 12).
- The outcome of the Washington Redskins football games could correctly predict the winner of this presidential election (cluster 13).
- Catholic consciences would recall the abortion issue in this election (cluster 14).
- American people were happy with the outcome (cluster 16).

- The voting machine worked well (cluster 17).
- President Bush would bring a new generation of conservative justices to the Supreme Court (cluster 18).
- The European people viewed the election differently (cluster 19).

Clusters: [2, 3, 7, 8, 15] referred to different types of political judgements and decisions. Clusters: [16, 20] referred to the outcome of the US presidential election 2004 and inauguration in 20 January 2005.

At the beginning, the US presidential election event triggered many questions and some responses and the outcome of the election occurred at the end. The election and bird flu events had the same characteristics, in the sense of triggering predictive assessments in a particular situation. This indicates the nature of the two events which have an inherent uncertainty about their final outcome, animal-human/human-human virus transmission for bird flu.

5.2. Discussion

Instead of operating at the document (i.e. RSS item) level, we operated at the level of features and selected a portion of significant terms. The terms formed a basis for further data processing. By using co-occurrence with significant terms, for each day we clustered all the terms which were contextually related. The generated clusters, however, were sometimes incorrect, as discussed in Section 4.3, and our approach did not achieve a very high level of success. Despite these weaknesses, our approach can be applied to generate overview timelines for major events, as described in Section 5.1. The case studies show that the clusters were genuinely related to the major events and could be used to form a narrative, albeit partial.

The human-based evaluation and analysis, as explained in Sections 4.2 and 5.1, are useful for determining the degree of the effectiveness of our approach from human interpreter's perspective. An automatic, metric-based evaluation, such as described in Ng and Han (2002) or Alexandrov et al. (2005), would have also been useful, if we would have had a set of predefined clusters as a gold standard, so that we could measure the cluster quality by measuring the number of overlaps between the predefined and automatically generated clusters. In our experimental setting, we did not have the predefined clusters with which our results could be compared. For these reasons, it was not possible for us to conduct an automatic evaluation.

There are some delays in identifying events, when compared with the other news media. An individual may use an RSS feed to express and post their comments and the propagation of news from a media source to the individual may take time. For example, the intention of Pope John Paul II to offer a special Mass on New Year's Eve for the tsunami victims was reported by Catholic World News (CWN) on 31/12/2005. Our cluster which is associated with the event was dated at 01/01/2005 (1 day later).

6. Conclusions and future work

Our method automatically produced clusters of terms from RSS feeds, which were assessed by human evaluators to see whether they appeared to signify a single news event. The low level of agreement among the 10 assessors ($\kappa_{\text{average}} = 0.36$) indicated the difficulty of the human task of reliably identifying an event from a small set of terms rather than problems with the clustering algorithm itself. The evaluation of 100 real clusters carried out by the assessors indicated that the average percentage of good clusters was 68.6%, which was much higher than the 16.7% for bad clusters. Thus, the method was clearly effective to some extent ($p < 0.01$). Our clustering approach, however, produces some incorrect clusters, as well as some clusters that it would be unreasonable to expect a non-expert to identify.

Despite these issues, our clustering approach can be applied to generate overview timelines for major events. The case studies show that the method can be applied to obtain coherent, human identifiable events and form a narrative, albeit partial. In future work, we hope to adopt our approach to fully automatically generate a short summaries of major events and their subsequent events from RSS feeds.

Acknowledgements

The work was supported by a European Union grant for activity code NEST-2003-Path-1. It is part of the Critical Events in Evolving Networks project (CREEN, contract 012684). We thank the reviewers for their helpful comments.

References

Adar, E., Zhang, L., Adamic, L. A., & Lukose, R. M. (2004). Implicit structure and the dynamic of blogspace. In *Proceedings of the 13th international WWW conference – workshop on weblogging ecosystem – aggregation, analysis and dynamics*.

Alexandrov, M., Gelbukh1, A., & Rosso, P. (2005). An approach to clustering abstracts. In *Proceedings of the 10th international conference on applications of natural language to information systems (NLDB 2005)* (pp. 275–285).

Allan, J., Carbonell, J., Doddington, G., Yamron, J., & Yang, Y. (1998). Topic detection and tracking pilot study: Final report. In *Proceedings of the DARPA broadcast news transcription and understanding workshop*.

Allan, J., Papka, R., & Lavrenko, V. (1998). On-line new event detection and tracking. In *Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval* (pp. 37–45).

Baeza-Yates, R., & Ribeiro-Netto, B. (1999). *Modern information retrieval* (1st ed.). ACM Press/Addison Wesley.

Belew, R. K. (2000). *Finding out about—a cognitive perspective on search engine technology and the WWW* (1st ed.). Cambridge University Press.

Brill, E. (1992). A simple rule-based part-of-speech tagger. In *Proceedings of the 3rd conference on applied natural language processing (ANLP 1992)* (pp. 152–155).

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37–46.

Eugenio, B. D., & Glass, M. (2004). The kappa statistics: A second look. *Computational Linguistics*, *30*(1), 95–101.

Glance, N. S., Hurst, M., & Tomokiyo, T. (2004). BlogPulse: Automated trend discovery for weblogs. In *Proceedings of the 13th international WWW conference – workshop on weblogging ecosystem – aggregation, analysis and dynamics*.

Gruhl, D., Guha, R., Liben-Nowell, D., & Tomkins, A. (2004). Information diffusion through blogspace. In *Proceedings of the 13th international WWW conference* (pp. 491–501).

Herring, S. C., Scheidt, L. A., Bonus, S., & Wright, E. (2004). Bridging the gap: A genre analysis of weblogs. In *Proceedings of the 37th Hawaii international conference on system sciences (HICSS-37)*.

Lin, J., & Halavais, A. (2004). Mapping the blogosphere in America. In *Proceedings of the 13th international WWW conference – workshop on weblogging ecosystem – aggregation, analysis and dynamics*.

MacKay, D. J. C. (2003). *Information theory, inference, and learning algorithms* (2nd ed.). Cambridge University Press.

Ng, R. T., & Han, J. (2002). CLARANS: A method for clustering objects for spatial data mining. *IEEE Transactions on Knowledge and Data Engineering*, *14*(5), 1003–1016.

NIST Speech Group. (2005). The topic detection and tracking phase 2 (TDT2) evaluation plan. [http://www.nist.gov/speech/tests/tdt/tdt98/ (accessed 15 June 2005)].

Prabowo, R., & Thelwall, M. (2006). A comparison of feature selection methods for an evolving RSS feed corpus. *IPM*, *42*(6), 1491–1512.

Ramshaw, L. A., & Marcus, M. P. (1995). Text chunking using transformation-based learning. In D. Yarovsky & K. Church (Eds.), *Proceedings of the 3rd workshop on very large corpora (VLC 1995)* (pp. 82–94).

Salton, G., & McGill, M. J. (1986). *Introduction to modern information retrieval* (1st ed.). McGraw-Hill, Inc.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, *34*(1), 1–47.

Shannon, C. E., & Weaver, W. (1963). *The mathematical theory of communication*. University of Illinois Press.

Siegel, S., & Castellan, J. N. J. (1988). *Nonparametric statistics for the behavioral sciences* (2nd ed.). McGraw-Hill.

Singhal, A., Buckley, C., & Mitra, M. (1996). Pivoted document length normalization. In H.-P. Frei, D. Harman, P. Schäuble, & R. Wilkinson (Eds.), *Proceedings of the 19th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 21–29).

Smith, D. A. (2002). Detecting and browsing events in unstructured text. In *Proceedings of the 25th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 73–80).

Stein, B., & Niggemann, O. (1999). On the nature of structure and its identification. In P. Widmayer, G. Neyer, & S. Eidenbenz (Eds.), *Proceedings of the 25th international workshop on graph-theoretic concepts in computer science* (pp. 122–134).

Swan, R., & Allan, J. (2000). Automatic generation of overview timelines. In E. Yannakoudakis, N. J. Belkin, M.-K. Leong, & P. Ingwersen (Eds.), *Proceedings of the 23rd annual international ACM SIGIR conference on research and development in information retrieval* (pp. 49–56).

Thelwall, M., Prabowo, R., & Fairclough, R. (2006). Are raw rss feeds suitable for broad issue scanning? A science concern case study. *JASIST*, *57*(12), 1644–1654.

Yang, Y., Pierce, T., & Carbonell, J. (1998). A study on retrospective and online event detection. In *Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval* (pp. 28–36).

Available online at www.sciencedirect.com
ScienceDirect
Journal of Informetrics 1 (2007) 145–154
Journal of INFORMETRICS
An International Journal
www.elsevier.com/locate/joi

Lifting the crown—citation *z*-score

Jonas Lundberg [a,b,*]

[a] *Medical Management Centre at the Department of Learning, Informatics, Management and Ethics (LIME), Karolinska Institutet, SE-17177 Stockholm, Sweden*
[b] *Strategy and Development Office, Karolinska Institutet, SE-17177 Stockholm, Sweden*

Received 31 August 2006; received in revised form 22 September 2006; accepted 22 September 2006

Abstract

Researchers worldwide are increasingly being assessed by the citation rates of their papers. These rates have potential impact on academic promotions and funding decisions. Currently there are several different ways that citation rates are being calculated, with the state of the art indicator being the *crown indicator*. This indicator has flaws and improvements could be considered. An *item oriented field normalized citation score average* ($\bar{c}_f$) is an incremental improvement as it differs from the crown indicator in so much as normalization takes place on the level of individual publication (or item) rather than on aggregated levels, and therefore assigns equal weight to each publication. The normalization on item level also makes it possible to calculate the second suggested indicator: total field normalized citation score (Σc_f). A more radical improvement (or complement) is suggested in the *item oriented field normalized logarithm-based citation z-score average* ($\bar{c}_{fz[ln]}$ or *citation z-score*). This indicator assigns equal weight to each included publication and takes the citation rate variability of different fields into account as well as the skewed distribution of citations over publications.

Even though the citation *z*-score could be considered a considerable improvement it should not be used as a sole indicator of research performance. Instead it should be used as one of many indicators as input for informed peer review.

Keywords: Citation; Indicator; Normalization; Research Assessment; *z*-score

1. Introduction

Researchers worldwide are increasingly being assessed by the citation rates of their papers. These rates have potential impact on academic promotions and funding decisions. Currently there are several different ways in which citation rates are being calculated, from basic calculations like raw citation counts and the h-index (Hirsch, 2005) to citation rates controlled for research field, publication year and document type. The reason for controlling for the first two factors is shown in Fig. 1. As can be seen in the figure articles in different research fields and from different publication years have varying average citation rates. Articles in *Cell Biology* journals on average receive about five times as many citations as do articles in *Crystallography* journals. The difference between fields is quite consistent over time. The older a publication is, the more likely it is that it has been cited (regardless of field). As seen in the figure, articles in the displayed areas on average are cited between six (*Biochemistry & Molecular Biology*) and eight (*Clinical Neurology*) times more often if they were published in 1998 than if they were published 6 years later.

* Correspondence address: Medical Management Centre at the Department of Learning, Informatics, Management and Ethics (LIME), Karolinska Institutet, SE-17177 Stockholm, Sweden
E-mail address: jonas.lundberg@ki.se.

1751-1577/$ – see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.joi.2006.09.007

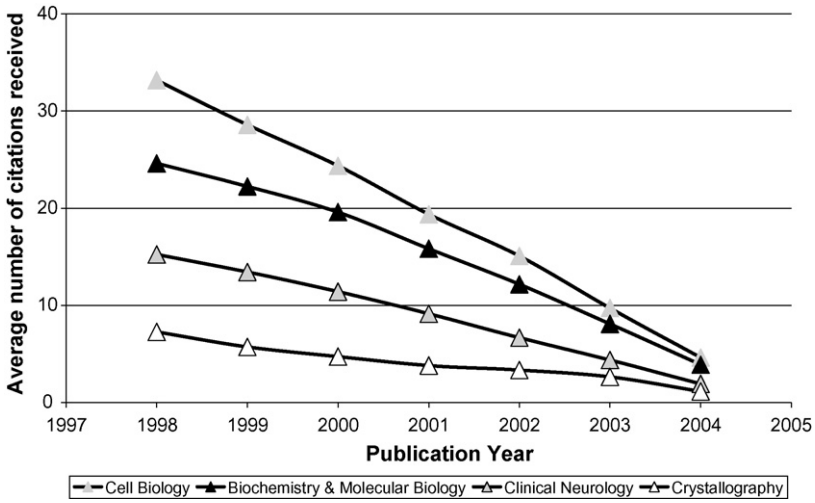


Fig. 1. Average citation rates 1998–2004 for articles in *Cell Biology*, *Biochemistry & Molecular Biology*, *Clinical Neurology and Crystallography*. Data from the citation indices produced by Thomson Scientific, self citations are included and whole counting is performed.

In Fig. 2 an example is shown of why it is important to control for document type. Publications classified as *Articles* receive about two fifths as many citations as publications classified as *Reviews* (35–40% between 1998 and 2004). *Letters* in turn receive one fifth as many citations as *Articles* (17–21% between 1998 and 2004). In summary, there are obvious differences in average citation rates for publications of different types, of different age, published in journals within different fields.

The first suggestions on how to control for these factors and calculate 'normalized' citation rates were made in the 1980s (Schubert, Glänzel, & Braun, 1983; Vinkler, 1986). In general, all proposed normalization methods are computed by dividing the actual number of received citations for a group of publications with the number of citations that could be expected for similar publications. Currently the state of the art indicator is the so called '*crown indicator*', developed at the Centre for Science and Technology Studies (CWTS) at Leiden University (Moed, Debruin, & Vanleeuwen, 1995). The indicator is calculated by dividing the average number of received citations for a group of publications with the average number that could be expected for publications of the same type, from the same year, published in journals within the same field. An example is shown in part I of Table 1.

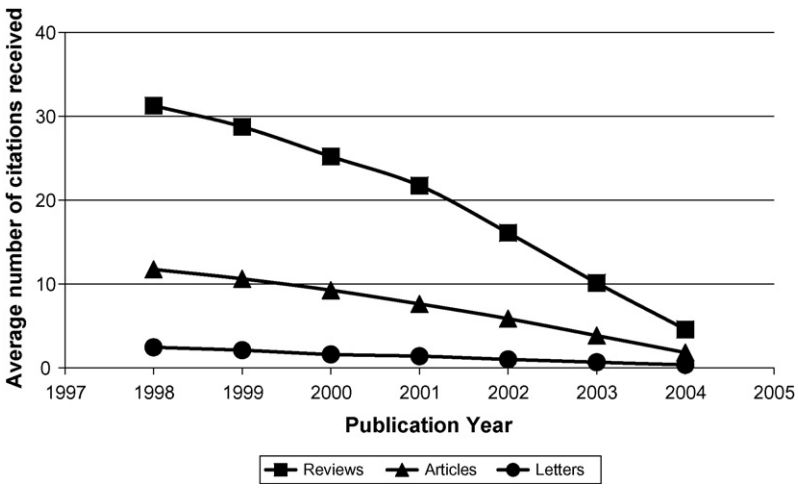


Fig. 2. Average citation rates for articles, letters and reviews in CI 1998–2004. Data from the citation indices produced by Thomson Scientific, self citations are included and whole counting is performed.

Table 1
Calculation of the 'crown indicator', item oriented field normalized citation score average, and item oriented field normalized logarithm-based citation z-score average

Article	Type	Year	Journal	Field	c	μ_f	σ_f	c_f	$\ln(c+1)$	$\mu_{fz[ln]}$	$\sigma_{fz[ln]}$	$c_{fz[ln]}$
A	Article	2003	*J. Appl. Crystallogr.*	Crystallography	12	2.6	22.1	4.6	2.6	0.9	0.8	2.2
B	Article	2003	*J. Cryst. Growth*	Crystallography	5	2.6	22.1	1.9	1.8	0.9	0.8	1.2
C	Review	2000	*Nat. Rev. Immunol.*	Immunology	66	33.4	76.1	2.0	4.2	2.7	1.3	1.1
D	Article	2000	*J. Immunol.*	Immunology	17	17.0	27.4	1.0	2.9	2.3	1.2	0.6
E	Letter	2001	*Am. J. Hum. Genet.*	Genetics & Heredity	17	9.7	32.2	1.8	2.9	1.6	1.2	1.1

c: Number of received citations; μ_f: the average value of citations to publications of the same type, published the same year in the same research area; σ_f: standard deviation for the average number of citations received for publications of the same type, from the same year, published in journals within the same field; c_f: field normalized citation score (c/μ_f); $\ln(c+1)$: logarithm of the number of received citations plus one; $\mu_{fz[ln]}$: the logarithm-based field citation score: the average value of the logarithmic number of citations (plus one) to publications of the same type, published the same year in the same research area; $\sigma_{fz[ln]}$: the standard deviation of the $\mu_{fz[ln]}$ distribution; $c_{fz[ln]}$: field normalized logarithm-based citation z-score;

(I) 'Crown indicator': $\left(\sum_{i=1}^{P} c_i / \sum_{i=1}^{P} [\mu_f]_i\right) = (12+5+66+17+17/2.6+2.6+33.4+17.0+9.7) \approx (117/65.3) \approx 1.8.$

(II) $\bar{c}_f$—item oriented field normalized citation score average: $(1/P)\sum_{i=1}^{P}(c_i/[\mu_f]_i) = (4.6+1.9+2.0+1.0+1.8/5) \approx 2.2.$

(III) $\bar{c}_{fz[ln]}$—item oriented field normalized logarithm-based citation z-score average: $(1/P)\sum_{i=1}^{P}(\ln(c_i+1) - [\mu_{f[ln]}]_i/[\sigma_{f[ln]}]_i) = (2.2+1.2+1.1+0.6+1.1/5) \approx 1.2.$

c_i: number of citations to publication i; P: the unit's number of publications; $[\mu_f]_i$: the average value of citations to publications of the same type, published the same year in the same research area as article i; $[\mu_{f[ln]}]_i$: the logarithm-based field citation score; the average value of the logarithmic number of citations (plus one) to publications of the same type, published the same year in the same research area as article i; $[\sigma_{f[ln]}]_i$: the standard deviation of the $[\mu_{f[ln]}]_i$ distribution.

First the average citation rate ($\bar{c}$ or CPP in CWTS terminology) for the studied five publications is calculated ((12 + 5 + 66 + 17 + 17)/5 = 23.4). Second the expected average citation rate for publications of the same type, from the same year, within the same field is calculated ($\bar{\mu}_f$ or FCSm in CWTS terminology) ((2.6 + 2.6 + 33.4 + 17.0 + 9.7)/5 = 13.1). Finally the crown indicator value is received by dividing the actual citation rate with the expected citation rate (23.4/13.1 = 1.8).

In this study one letter acronyms are used in indicator calculations instead of the multi-letter acronyms suggested earlier (e.g. $\bar{c}$ instead of CPP for average number of citations per publication, or $\bar{\mu}_f$ instead of FCSm for average citation rate for a specific field, document type and publishing year). The use of one letter acronyms makes it possible to write down calculations more clearly.

2. Data sources

The figures in this study are based on data from the citation indices produced by Thomson Scientific.[1] These include the Science Citation Index (SCI), Social Science Citation Index (SSCI) and the Arts & Humantities Citation Index (AHCI). They are jointly referred to as the Thomson Scientific Citation Indices (CI). CI data has been imported into a *MySQL* database. The figures were created using Microsoft® Office Excel 2003. Self citations were included in all figures and whole counting was performed (not fractional counting). It should be noted that the suggested normalized citation rates of individual publications easily could be calculated and published by Thomson Scientific on their Web of Science®. Until this occurs the calculation of the suggested indicators requires direct access to world-wide citation data.[2]

3. Lifting the crown

It can be argued that the crown indicator has flaws. The first is that citation rates are not normalized on the level of individual publications, but on a higher aggregation level where the average citation rate of a researcher, group or department is compared to the average citation rate of the fields in which the researcher or group has published. This way of calculating gives more weight to older publications (particularly reviews), published in fields with dense citation traffic. In order to give each publication equal weight the normalization should take place on the level of the individual publication. The calculation of such an *item oriented field normalized citation score average* – $\bar{c}_f$– is shown in part II of Table 1. Here, instead of first calculating the actual average citation rate, and then divide that with the average expected citation rate, each publication (or item) is normalized individually (hence 'item oriented'). For example, article A has received 12 citations (c). The average number of citations that similar publications (*Articles* from 2003 published in *Crystallography* journals) have received is 2.6 (μ_f). The field normalized citation score (c_f) is thus $c/\mu_f = 12/2.6 = 4.6$. This procedure is repeated for each of the publications that the unit one intends to assess has published. In the example one continues with B (c_f: 1.9), C (2.0), D (1.0) and E (1.8). In the final step, the average for the field normalized citation scores is calculated ($\bar{c}_f = (4.6 + 1.9 + 2.0 + 1.0 + 1.8)/5 = 2.2$).

The field normalized citation scores can also be summed in order to calculate a total field normalized citation score (Σc_f) for a research group, university or country.

4. Citation z-score

The distribution of citations over publications differs between 'normalization groups' (publications of a specific type, within a specific research field, published a specific year). Therefore, does not only the average citation rate differ, but also the standard deviation. It could thus be argued that using a z-score in the normalization procedure would be more appropriate. A z-score expresses how far a value is from the population mean, and expresses this difference in terms of the number of standard deviations by which it differs (Kirkwood & Sterne, 2003). A second issue that needs to

[1] Certain data included herein is derived from the 1995 to 2005 Science Citation Index Expanded, Social Sciences Citation Index and, Arts & Humanities Citation Index Tagged Data prepared by the Thomson Scientific Inc. (TS), Philadelphia, Pennsylvania, USA: ©Copyright Thomson Scientific Inc® 2006. All rights reserved.

[2] For more information on how Karolinska Institutet has developed its bibliometric system please contact the author.

be dealt with is that the distribution of citations over publications is highly skewed. The skewed distribution has been seen for fields and journals, as well as for individual scientists (Seglen, 1992). It could thus be an alternative to use the geometric mean or median value as comparison when calculating normalized citation rates. Another alternative is to make normalizations using logarithmically transformed citation rates.

Combining the observations above, it could be argued that it would be appropriate to use a logarithmic *citation z-score* as a complementary indicator to the item oriented field normalized citation score average. A citation *z*-score would compare the logarithm of the number of citations that a publication has received with the mean and standard deviation for the logarithms of the citation rates for all the corresponding reference publications.

When calculating the proposed indicator one should add one to the number of citations for each publication. This is necessary since it is not possible to calculate the logarithm of zero. For example, a review article published in 2000 in *Nature Reviews Immunology* (Article C in Table 1) has received 66 citations. The natural logarithm of this value plus one ($\ln(67) = 4.2$) would then be compared with the average number (2.7) and standard deviation (1.3) of the natural logarithms of citation rates (plus one) of all reviews from 2003 in immunology. The citation *z*-score for this article is then $(4.2 - 2.7)/1.3 \approx 1.1$. Observe that the comparison is made with *average of the logarithms* of the number of citations received by comparable items and *not* with the *logarithm of the average* number of citations received by comparable items. The bibliometric indicator for a research group, department or university is then the *item oriented field normalized logarithm-based citation z-score average* ($\bar{c}_{fz[\ln]}$—or '*citation z-score*'). An example of the calculation is shown in part III of Table 1. The citation *z*-score would in this case be $(2.2 + 1.2 + 1.1 + 0.6 + 1.1)/5 = 1.2$. The publications in the example are thus, after logarithmic transformation, on average cited 1.2 standard deviations above the world average for publications of the same type, from the same year, published in journals belonging to the same subject category.

5. Examples

5.1. Distribution of publications over $c_{fz[ln]}$

Between 1998 and 2003, 60,082 publications with at least one co-author affiliated with an organisation in Sweden, were published in journal belonging to 57 life science related areas.[3] 49,490 were articles, letters or reviews belonging to normalization groups that meet the inclusion criteria that the average citation rate should be at least one citation and there should be at least 100 publication of the same type, from same year, published in journals within the same subject category. The Swedish life science publications on average have been cited 26% above the world average. After logarithmic transformation, the average publication has received citations 0.23 standard deviations above the world average (mean: 0.23; median: 0.26; max 5.9; min −2.51). The distribution of publications over $c_{fz[\ln]}$ is shown in Fig. 3. As can be seen in the figure the distribution is approximately normal with a slight positive skew (skewness: 0.03).

Similarly to the shown distribution of publications over $c_{fz[\ln]}$ for a country (as shown in Fig. 3) the distribution can be shown on meso (university or department) and micro level (individual researcher or research group). An example of how the citation rates for publications by research groups builds up the distribution for a whole department is shown in Fig. 4. The figure is based on 453 publications by 20 research groups constituting a department at Karolinska Institutet (KI), published between 1998 and 2003. The research groups have published between 3 and 70 publications during the time period. The distribution of publications over categories of $c_{fz[\ln]}$ for each of the 20 research groups is represented by a line. The lines are placed on top of each other, giving the distribution for the department as a whole.

[3] Acoustics, Anesthesiology, Behavioral Sciences, Biochemistry & Molecular Biology, Biophysics, Biotechnology & Applied Microbiology, Cardiac & Cardiovascular System, Cell Biology, Clinical Neurology, Critical Care Medicine, Dentistry, Oral Surgery & Medicine, Dermatology & Venereal Diseases, Education, Special, Endocrinology & Metabolism, Engineering, Biomedical, Gastroenterology & Hepatology, Genetics & Heredity, Geriatrics & Gerontology, Gerontology, Health Care Sciences & Services, Hematology, Immunology, Infectious Diseases, Language & Linguistics Theory, Medical Informatics, Medicine, General & Internal, Medicine, Legal, Medicine, Research & Experimental, Microbiology, Multidisciplinary Sciences, Neurosciences, Nursing, Obstetrics & Gynecology, Oncology, Ophthalmology, Orthopedics, Otorhinolaryngology, Pathology, Pediatrics, Peripheral Vascular Diseases, Pharmacology & Pharmacy, Philosophy, Psychiatry, Psychology, Clinical, Psychology, Development, Public, Environmental & Occupational Health, Radiology, Nuclear Medicine & Medical Imaging, Rehabilitation, Reproductive Biology, Respiratory System, Rheumatology, Social Issues, Social Sciences, Interdisciplinary, Sport Sciences, Surgery, Transplantation, Urology & Nephrology.

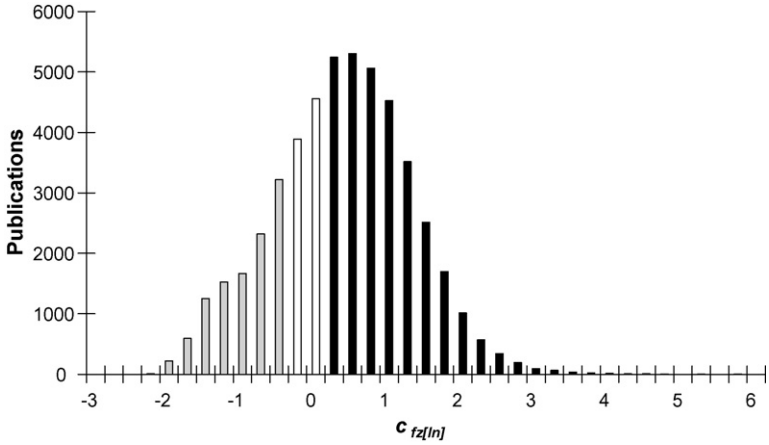


Fig. 3. Distribution of publications over classes of $c_{fz[ln]}$. The figure is based on 49,490 Swedish publications in 57 life science related areas, published between 1998 and 2003. Bars including publications with $c_{fz[ln]} < -0.25$ are coloured grey, $c_{fz[ln]}$ between −0.25 and +0.25 white, and $c_{fz[ln]} > +0.25$ black. Data from the citation indices produced by Thomson Scientific, self citations are included and whole counting is performed.

The distribution of publications over $c_{fz[ln]}$ can be used for comparisons between different aggregation levels, for example a department, a university and a country. In Fig. 5 the relative distribution of publications over $c_{fz[ln]}$ for a department at KI is compared with the distribution for KI as a whole, and the distribution of all Swedish publications within the 57 life science related research areas (1998–2003). In the example 61% of the Swedish life science publications have a $c_{fz[ln]}$ above zero, while KI as a whole has 66% and the selected department 69%. If instead looking at the share of publications with a value above one, KI as a whole has a higher share (24%) than the department (22%).

5.2. Distribution of units over citation z-score

On higher aggregation levels distribution of citation z-scores also approaches normal distribution. An example is shown in Fig. 6. The figure show data for 334 research units at KI that have published at least 30 publications between 1995 and 2004 (mean $\bar{c}_{fz[ln]}$ 0.37; median 0.36; min −0.39; max 1.32). Ninety percent of the units have citation z-scores above 0, and 3% have a value more than one standard deviation above world average.

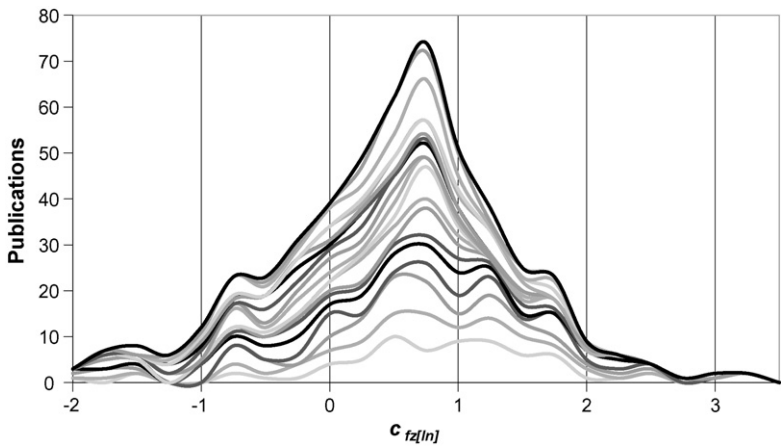


Fig. 4. Distribution of publications over classes of $c_{fz[ln]}$. The figure is based on 453 publications by the 20 research groups at a department at Karolinska Institutet, published between 1998 and 2003. Each line represent the publications by a research group. The lines are placed on top of each other. Data from the citation indices produced by Thomson Scientific, self citations are included and whole counting is performed.

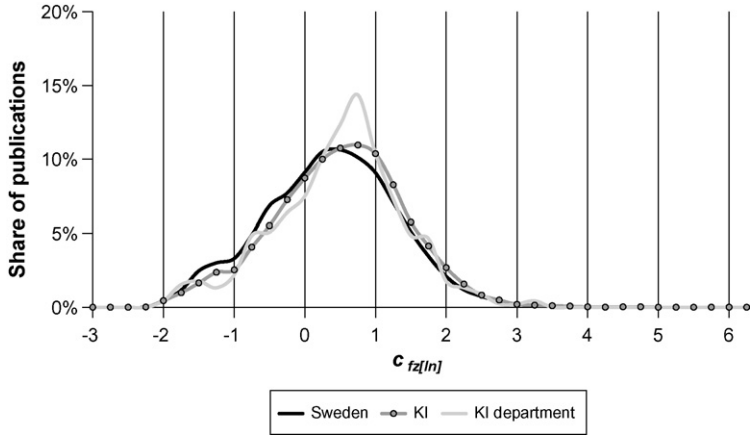


Fig. 5. Relative distribution of publications over classes of $c_{fz[ln]}$ for a department at KI, KI as a whole, and all Swedish publications within 57 research areas (1998–2003).

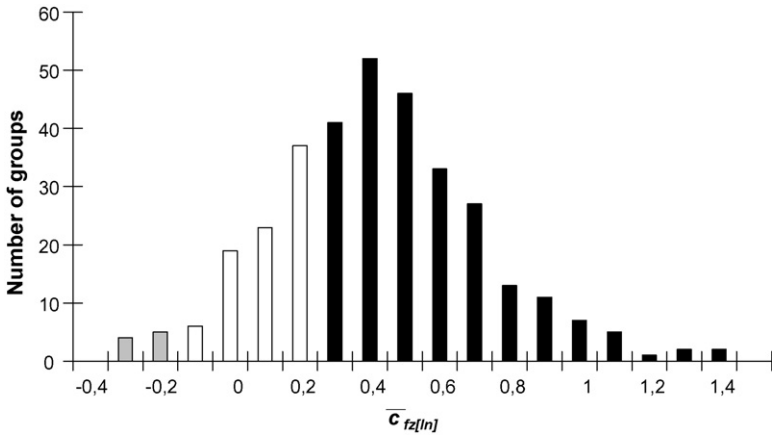


Fig. 6. Distribution of research units over classes of citation z-scores. The figure is based on data for 334 research unites at KI (at least 30 publications 1995–2004). Bars including groups with citation z-score <-0.2 are coloured grey, citation z-score -0.2 and $+0.2$ white, and citation z-score $>+0.2$ black. Data from the citation indices produced by Thomson Scientific, self citations are included and whole counting is performed.

5.3. Development on country level

How Σc_f and $\bar{c}_{fz[ln]}$ can be used on macro level is shown in figure Figs. 7 and 8. In Fig. 7 indicator values are shown for life science publications published between 1998 and 2003 by researchers affiliated with organisations in USA, France, Germany, Russia and China. As can be seen in the figure, the average citation rate is highest for USA, but both Germany and France are catching up. A faster increase can be seen for China and Russia, but from a much lower level.

When instead looking at the total field normalized citation score (Σc_f) size becomes a factor (Fig. 8). The indicator value for the United States is six to nine times larger than for Germany and France. The development in China is apparent in the figure, going from about 1% of USA level in 1998 to 4% in 2003.

6. Discussion: an improvement but still many caveats

In this study three new indicators are suggested. The indicators build on earlier research which has shown how citation rates could be normalized for research field, publication year and document type (Moed et al., 1995; Schubert et al., 1983; Vinkler, 1986). The *item oriented field normalized citation score average* ($\bar{c}_f$) is an incremental improvement of the current state of the art indicator (the crown indicator). It differs from the crown indicator in so much as it assigns equal weight to each included publication. That normalization takes place on the level of individual item also makes

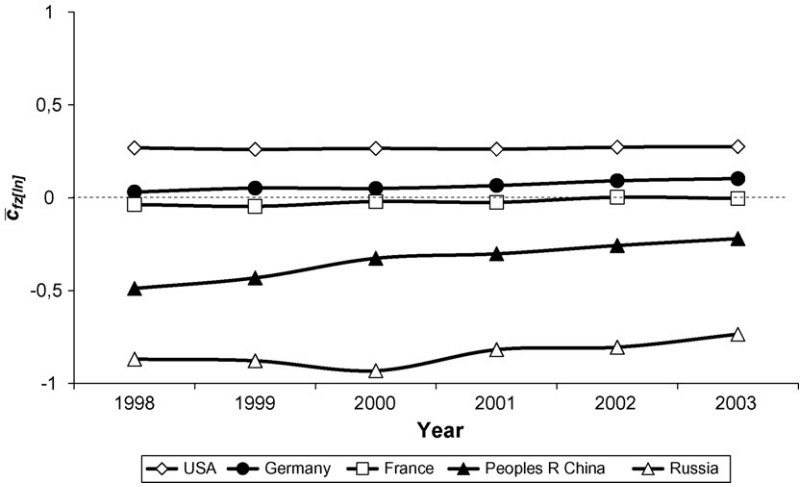
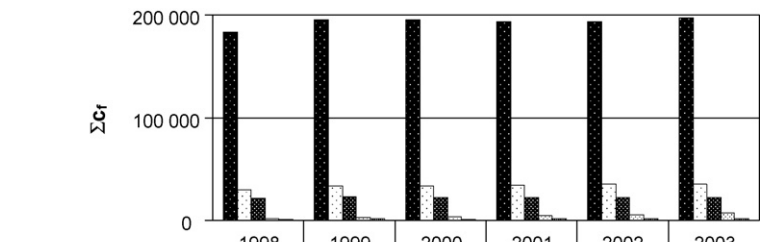


Fig. 7. Item oriented field normalized logarithm-based citation z-score average 1998–2003 for 1,416,912 publications in 57 life science related subject categories, co-authored by researchers affiliated with organisations in USA, Germany, France, China and Russia. Data from the citation indices produced by Thomson Scientific, self citations are included and whole counting is performed.

it possible to calculate total field normalized citation score (Σc_f) (example in Fig. 8). The more radical improvement (or complement) is the *item oriented field normalized logarithm-based citation z-score average* ($\bar{c}_{fz[ln]}$ or *citation z-score*). This indicator assigns equal weight to each included publication and takes the citation rate variability of different fields as well as the skewed distribution of citations over publications into account. In comparison with the crown indicator, which distribution starts to approach normal on aggregated levels (such as distribution of research groups over classes of crown indicator values), the distribution of field normalized logarithm-based citation z-scores ($c_{fz[ln]}$) over publications starts to approach normal distribution already *within* low aggregation levels such as research groups (Fig. 4). $c_{fz[ln]}$ values are approximately normally distributed *within* department or university level (Fig. 5). On aggregated levels the citation z-score is approximately normally distributed on the level of research groups (Fig. 6). Since the citation z-score is based on logarithmic transformations of citation rates, extreme values have less impact on this indicator. This is both the strength and the weakness of the indicator since one is often interested in just those 'extreme cases" when assessing research (Tijssen, Visser, & van Leeuwen, 2002). The citation z-score thus provides a complementary view to other indicators which are more heavily influenced by extreme citation rates (e.g. $\bar{c}_f$) or solely



	1998	1999	2000	2001	2002	2003
USA	183 208	195 151	194 999	193 163	193 781	197 374
Germany	29 889	33 385	33 630	34 453	35 268	35 202
France	21 670	22 829	22 281	22 765	22 452	22 149
Peoples R China	1 942	2 602	3 827	4 755	5 580	7 704
Russia	1 368	1 426	1 343	1 460	1 526	1 577

Year

Fig. 8. The total field normalized citation score 1998–2003 for 1,416,912 publications in 57 life science related subject categories, co-authored by researchers affiliated with organisations in USA, Germany, France, China and Russia. Data from the citation indices produced by Thomson Scientific, self citations are included and whole counting is performed.

concerns very highly cited publications (e.g. share of publications in top 1% of the citation distribution). It therefore seems appropriate to provide information on both types of indicators as input to informed peer review.

When calculating bibliometric indicators there are some important issues to consider. First one could discuss whether to control for research field or consider research fields with high average citation rates as being of higher quality. One might argue that research performed within fields with high average citation rates, such as cell biology or biochemistry, on average is of higher quality than research within the fields of nursing or clinical neurology. One should then remember that the clinical neurologists easily could increase their average citation rate (and thus their 'quality' as measured by raw average citation rate) by starting to write longer reference lists. This would of course not increase the quality of research in clinical neurology.

Secondly, one should consider the assignment of articles to research fields based on the subject categories of journals (Glänzel & Schubert, 2003). Currently the only classification scheme that is readily available for evaluative purposes across scientific disciplines is the ISI subject categories of journals produced by Thomson Scientific. An issue with using this method is that it has been shown that publications within one field are often published in journals that are categorised as belonging to another field (e.g. Lewison, 1996; Lundberg, Fransson, Brommels, Skar, & Lundkvist, 2006; Ugolini, Casilli, & Mela, 2002). Since the categorization is made on journal level, a related issue concerns articles published in multidisciplinary journals such as *Science*. Articles in these journals could for example be normalized using the average citation rate of all other articles published in journals within the 'multidisciplinary-field', or they could individually be assigned to its respective research field. The latter is time-consuming but the number of multidisciplinary journals is limited, so using a collaborative effort the task of manually categorizing these articles should not be overwhelming. Other classification schemes are being developed, and hopefully it should not take long before there are standardized classification schemes on the level of individual publications available across disciplines similar to ones that are available today within single fields, e.g. Medical Subject Headings (MeSH) for medicine.

A third issue is whether to control for document type. That review and 'normal' journal articles differ in average citation rates as well as in type of research is uncontroversial. A more contentious issue is whether to control for the document type *letter*, as these publications could be seen simply as short normal journal articles. Their average citation rate is, however, much lower than normal articles'. In the suggested indicators, letters and normal articles are separated in the normalization procedure and the choice of what document types to include is made when calculating the final indicator of a researcher, group or department. One could simply choose to omit letters from the calculation of the final indicator values or display indicators for each publication type separately.

Fourth, two of the indicators described here deal with the quality aspect of research performance. In any research assessment quantity also needs to be considered. Calculation of composite indicators has been suggested by, for example, Lindsey (1978). The $\sum c_f$ gives an indication of the total international impact that an assessed group, organisation or country has had. It should be noted that extreme 'quality' indicator values (both low and high) are to be expected for units with a low number of publications. This is of importance when using the suggested indicators for comparing units of different size (for example small and large research groups within a university). A related issue that needs to be dealt with is the decision whether to conduct whole or fractional counting of citation rates. Should each contributor receive full credit for a paper, or should the credit be distributed according to some formula?

Fifth, a limitation when calculating the suggested normalized citation rates is that the average citation rate (and standard deviation) of a field cannot be zero. It could even be argued that the lower limit for the average citation rate should be at least one citation in order for the calculations of indicators to be valid. A second lower limit can also be (arbitrarily) set for the minimum number of publications in the normalization group—for example that there should be at least 100 publication of the same type, from same year, published in journals within the same subject category in order for a publication to be included in the normalization procedure.

The bibliometric community owes the rest of the scientific community not only to deal with the issues mentioned here, but also to follow Professor Wolfgang Glänzel's 10 year old recommendations; to work on definitions, reproducibility, validation and compatibility (Glänzel, 1996) and to continue work on new indicators that can contribute to the development of science world wide.

7. Conclusion

An *item oriented field normalized logarithm-based citation z-score average* ($\bar{c}_{fz[ln]}$) is an improvement of available bibliometric indicators as it assigns equal weight to each included publication, takes the citation rate variability of

different fields into account and also considers the skewed distribution of citations over publications. Still, it should not be used as a sole indicator of research performance but rather as one of many indicators used as input for informed peer review.

Acknowledgements

The author would like to acknowledge the valuable suggestions and feedback provided by Catharina Rehn and Ulf Kronman at Karolinska Institutet University Library, and by two anonymous reviewers.

References

Glänzel, W. (1996). The need for standards in bibliometric research and technology. *Scientometrics*, *35*(2), 167–176.

Glänzel, W., & Schubert, A. (2003). A new classification scheme of science fields and subfields designed for scientometric evaluation purposes. *Scientometrics*, *56*(3), 357–367.

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*(46), 16569–16572.

Kirkwood, B. R., & Sterne, J. A. C. (2003). *Essential medical statistics* (2nd ed.). Malden, Mass: Blackwell Science.

Lewison, G. (1996). The definition of biomedical research subfields with title keywords and application to the analysis of research outputs. *Research Evaluation*, *6*(25), 25–36.

Lindsey, D. (1978). The corrected quality ratio: A composite index of scientific contribution to knowledge. *Social Studies of Science*, *8*(3), 349–354.

Lundberg, J., Fransson, A., Brommels, M., Skar, J., & Lundkvist, I. (2006). Is it better or just the same? Article identification strategies impact bibliometric assessments. *Scientometrics*, *66*(1), 183–197.

Moed, H. F., Debruin, R. E., & Vanleeuwen, T. N. (1995). New bibliometric tools for the assessment of National Research Performance—Database description, overview of indicators and first applications. *Scientometrics*, *33*(3), 381–422.

Schubert, A., Glänzel, W., & Braun, T. (1983). Relative citation rate: A new indicator for measuring the impact of publications. In *Paper presented at the first national conference with international participation on scientometrics and linguistics of scientific text*.

Seglen, P. O. (1992). The skewness of science. *Journal of the American Society for Information Science*, *43*(9), 628–638.

Tijssen, R. J. W., Visser, M. S., & van Leeuwen, T. N. (2002). Benchmarking international scientific excellence: Are highly cited research papers an appropriate frame of reference? *Scientometrics*, *54*(3), 381–397.

Ugolini, D., Casilli, C., & Mela, G. S. (2002). Assessing oncological productivity: is one method sufficient? *European Journal of Cancer*, *38*(8), 1121–1125.

Vinkler, P. (1986). Evaluation of some methods for the relative assessment of scientific publications. *Scientometrics*, *10*(3–4), 157–177.

Available online at www.sciencedirect.com
ScienceDirect
Journal of Informetrics 1 (2007) 155–160
Journal of INFORMETRICS
An International Journal
www.elsevier.com/locate/joi

On the behavior of journal impact factor rank-order distribution

R. Mansilla [a], E. Köppen [a], G. Cocho [b], P. Miramontes [c,*]

[a] *Centro de Investigaciones Interdisciplinarias en Ciencias y Humanidades, Universidad Nacional Autónoma de México, México 04510, D.F., Mexico*
[b] *Instituto de Física, Universidad Nacional Autónoma de México, México 04510, D.F., Mexico*
[c] *Facultad de Ciencias, Universidad Nacional Autónoma de México, México 04510, D.F., Mexico*

Received 25 September 2006; received in revised form 30 December 2006; accepted 4 January 2007

Abstract

An empirical law for the rank-order behavior of journal impact factors is found. Using an extensive data base on impact factors including journals on education, agrosciences, geosciences, mathematics, chemistry, medicine, engineering, physics, biosciences and environmental, computer and material sciences, we have found extremely good fittings outperforming other rank-order models. Based in our results, we propose a two-exponent Lotkaian Informetrics. Some extensions to other areas of knowledge are discussed.

Keywords: Zipf's law; Lotkaian Informetrics; Power laws; Impact factors

1. Introduction

Quantitative studies in linguistics have a long lineage. Due to the extreme complexity of languages, these studies have been mainly based on statistical properties of words in literary corpora. Outstanding early examples of these studies are Estoup (1916), Dewey (1923) and Condon (1928). However, the most influential contribution on this topic is by Zipf (1949). In his work it appears what is today known as Zipf's law which can be formulated as follows: let $f(r)$, $r = 1, \ldots, N$, be the relative frequency of the words (the number of times a word appears divided by the total amount of words) in a text in decreasing order. Then Zipf's law states that:

$$f(r) = \frac{K}{r^{\alpha}}. \tag{1}$$

In this case, the items are words taken from a given text, the most abundant word takes the first place ($r = 1$), the second one takes the following place ($r = 2$) and so on. The fact that the mathematical expression of the law is a negative exponent power law implies that the law is a straight line with negative slope α when plotted in log–log scales. K is a proportionality constant with no phenomenological interest. This empirical law has found applications in a wide range of natural and human phenomena (Li, 2003). The case when $\alpha \simeq 1$ is of particular interest because it implies self-similarity.

* Corresponding author. Fax: +52 55 5622 4859.
E-mail address: pmv@fciencias.unam.mx (P. Miramontes).

1751-1577/$ – see front matter
doi:10.1016/j.joi.2007.01.001

The exact mechanism behind Zipf's law still remains a mystery so far. However, it is important to remark that the presence of power laws implies in general that the underlying mechanism is neither stochastic or regular. Power laws are the signature of correlated (colored) noise possibly indicating an "edge of chaos" dynamics (Langton, 1990) and all the rich phenomena associated (McElvey, 2001) or could be as well a clue to self-organized criticality (Bak, Tang, & Wiesenfeld, 1987).

The main drawback of Zipf's law was the bad fitting at very high and very low frequencies in the word counting problem. An improvement over the Zipf's law was proposed by Mandelbrot (1954):

$$f(r) = \left[\frac{N+\rho}{r+\rho}\right]^{1+\epsilon}, \tag{2}$$

where N is the number of different words in the text and ρ, ϵ are parameters to be adjusted.

Zipf's law is a special case of Mandelbrot's. This fact, along with a complete discussion of the role of power laws in the field of Informetrics can be found in Egghe (2005).

Recently it has been reported (Le Quan, Sicilia-García, Ming, & Smith, 2002) that what Zipf found is valid for small corpora (for the size of the texts that were analyzable at that time), and that today that the computer allows the analysis of huge texts, the log–log plot shows a clear downwards bending tail instead of the predicted straight line.

Scientific productivity is another topic where the first studies date back almost a century with the works of Dresden (1922) and Lotka (1926). The law of Lotka has the same mathematical form of Eq. (1) but he already introduced bibliometric variables by using r as contributors or authors of a given paper and $f(r)$ as articles or papers themselves. Since Lotka, it is common to call "sources" the independent variable and "item" the dependent one. This way, Lotka's law states that the number of items is a power law of the sources. The branch of Informetrics related to the study of power laws is called Lotkaian Informetrics (Egghe, 2005).

Informetrics mainly deals with the relationships between sources and items. It is normal to find the pairs authors–journals or journals–bibliographies as sources and items. In this paper we explore the possibility of extending the Lotkaian Informetrics to the realm of journal impact factors (JIFs). We show as well that the rank-order JIFs plots deviate from a traditional Lotkaian equation and propose an extension to what it could be called two-exponent Lotkaian laws.

2. Impact factors

Impact factor is a measure of the frequency with which the "average article" in a journal has been cited in a particular year or period (Garfield, 1994), it is calculated "by dividing the number of times a journal has been cited by the number of articles it has published during some specific period of time. The journal impact factor will thus reflect an average citation rate per published article" (Garfield, 1955). The impact factor of journals is an attempt to evaluate the knowledge production published among different journals of a given field. Mainly covered by the Science Citation Index database, it is published annually since 1975 in the Journal Citation Reports.

JIFs has been the target of many criticisms (Soegler, 1997; Fröhlich, 1996) and there is a debate about its usage as a tool to evaluate research. Even the influential journal Nature states that the JIFs figures should be handled carefully (Nature, 2005). Regardless its pros and cons, the fact is that it is an every day measure of the importance of a journal and it is worldwide used (de Marchi & Rocchi, 2001).

While keeping a skeptical attitude towards the use of the JIFs to evaluate scientific research, it should be recognized that it is an outcome of the process of publication and it has became by itself a subject of scientific study.

Rank-order distribution of JIFs attracted the attention of Lavalette who (mentioned in Popescu, 2003) proposed the following law:

$$f(r) = K\left[\frac{N+1-r}{r}\right]^{b} \tag{3}$$

where N is the number of journals, r the ranking number, $f(r)$ the impact factor, and b is a parameter to be fitted.

In the next section we propose a law that outperforms Lavalette's (see Section 5).

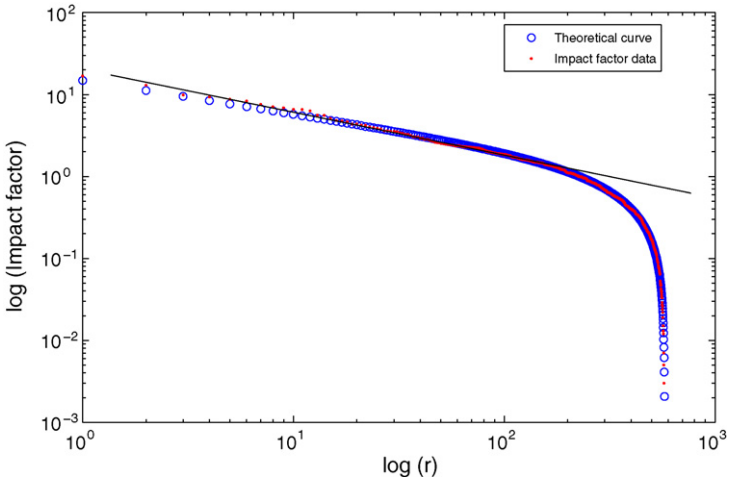


Fig. 1. Log–log rank-order plot of the impact factor data for physics journals. Notice the drop of the tail of the curve (see text).

Table 1

Scientific field	k	b	a	R^2
Physics	0.0273	0.991	0.4058	0.9999
Mathematics	0.0437	0.676	0.2622	0.9999
Computer science	0.0066	1.0626	0.2840	0.9999
Agroscience	0.0070	0.9597	0.2210	0.9999
Environmental science	0.0358	0.9357	0.2781	0.9800
Biosciences	0.0304	1.0161	0.5140	0.9999
Chemistry	0.0549	0.9733	0.4560	0.9999
Engineering	0.0033	1.0472	0.3522	0.9999
Geosciences	0.0463	0.8739	0.3505	0.9999
Material science	0.0408	0.9072	0.4477	0.9999
Medicine	0.0819	0.7735	0.4307	0.9999
Education	0.0819	0.7735	0.4307	0.9999

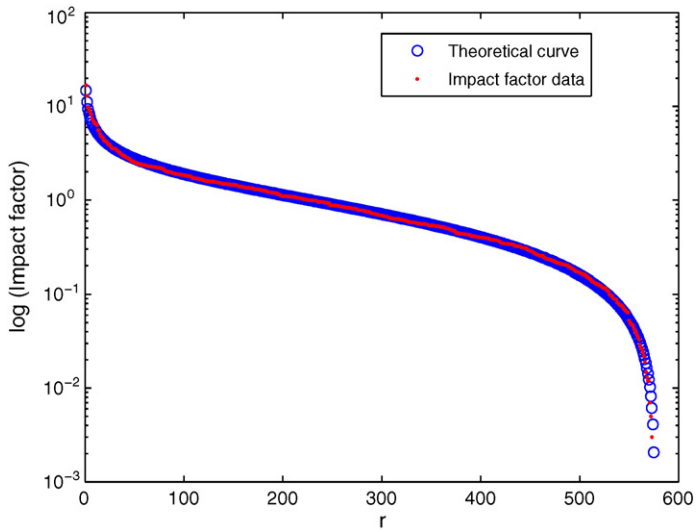


Fig. 2. Semi-log impact factor rank-order distribution for physics journals. Solid circles represent raw data. Hollow circles are the data evaluated in Eq. (4).

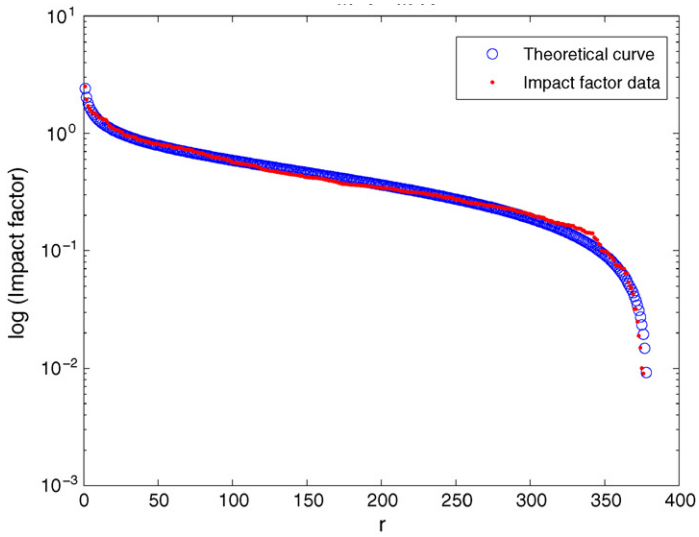


Fig. 3. Semi-log impact factor rank-order distribution for mathematics journals. Solid circles represent raw data. Hollow circles are the data evaluated in Eq. (4).

3. Analytical expression of the law

Fig. 1 shows the log–log plot of the IF of a randomly taken field from Popescu's database (2003).

It is evident that it is not a power law (the straight line was drawn as a reference) because of the bending tail in the right side of the plot. This fact motivated us to propose a beta-like function:

$$f(r) = K\frac{(N+1-r)^b}{r^a} \tag{4}$$

$f(r)$, $r = 1\ldots, N$ represents the rank-order impact factors; K, a and b are three parameters to fit. K is a meaningless scaling factor. Notice that when $b = 0$ this equation becomes Lotka's law.

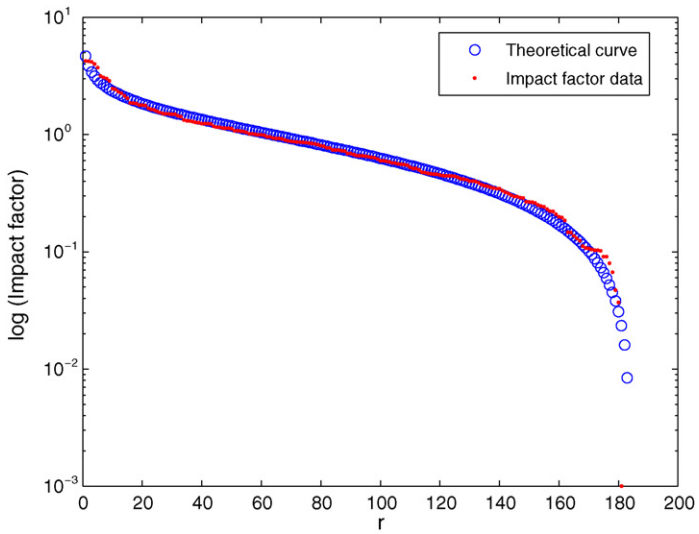


Fig. 4. Semi-log impact factor rank-order distribution for environmental sciences. Solid circles represent raw data. Hollow circles are the data evaluated in Eq. (4).

4. Results

For every set of data, we find the parameters values using the linear least squares method after transforming the coordinates to the logarithmic variable:

$$\log(f(r)) = \log(K) + b\,\log(N + 1 - r) - a\,\log(r) \tag{5}$$

Table 1 shows the values of K, b and a, as well as the coefficient of regression r^2 for impact factors of 12 disciplines. In Figs. 2–4 , the impact factors data as well as our theoretical curve for the fields of physical, mathematical and environmental sciences are shown. We used semilog plots because they are more natural when the abscissa is a rank-order variable.

The quality of the fitting is remarkable. Please notice from the analysis of Eq. (4) that the parameter a is more influential for small values of r. This fact means that for low values of r the phenomenon is nearly Lotkaian but this property is lost as the abscissae increase.

5. Concluding remarks

We have shown the excellent agreement of the data with our model. The quality of the fitting is superior to the proposal of Lavalette. From the comparison of Eqs. (3) and (4), it follows that Lavalette's law is a particular case of ours when $a = b$. Unfortunately, it is not possible to discuss the rationale behind Lavalette's law because the original paper is not available and all we know about it is a mention in Popescu's paper (2003).

The underlying proposed mechanism yielding the above-discussed behaviors often assumes a kind of "biological evolution form". For instance, Yule (1924) working in a model suggested by Willis (1922) managed to prove that assuming a single ancestral specie and probabilities of mutation and duplication a power law behavior is obtained. Expansion-modification systems proposed by Li (1991), which take into account the basic features of DNA mutation processes (Mansilla & Cocho, 2000), are also able to predict this behavior.

When discussing journal impact factors, a balance between the importance to the researchers of publish their work in high ranked journal, the difficulties associated with doing this and the increase of impact received by journals with high impact factor, seems to create a "rich gets richer" (the "Matthew Effect", see Merton, 1968; Egghe & Rousseau, 1990) mechanism also observed in the dynamics of complex networks (Barabasi, 2002). More than 49 years ago, Simon (1955, 1957) proposed a model which produces similar distributions. It is also interesting to notice that the bending of the tail of JIFs rank-order distribution means that after a critical zone of JIFs values is smooth thus discarding the possibility of the existence of multifractality.

Power-laws seem to be ubiquitous in physics, biology, geography, economics, linguistics, etc. (see Li, 2003). We consider "linguistic studies" not only those related with natural languages but also arbitrary languages over abstract finite alphabets. When the number of possible "words" is large, as it is the case for natural languages, it is expected to have a good fitness with a one-parameter power law. However, when the number of words is rather small, as it is the case of programming languages, one-exponent power laws absolutely fails and more parameters are necessary for a suitable fit. New elements to this considerations have been given by Le Quan et al. (2002). They showed that there is a serious deviation when the size of the sample is huge.

We expect that the increase in computing power will show that the deviation of Zipf's and Lotka's laws is a generic phenomenon. Then, a two-exponent Lotkaian and Zipfian Informetrics and linguistics should be welcome.

Acknowledgements

This work has been partially supported by the UNAM-PAPIIT grant IN-111003. The authors thank the sound comments of two anonymous referees.

References

Bak, P., Tang, C., & Wiesenfeld, K. (1987). Self-organized criticality: An explanation of $1/f$ noise. *Physical Review Letters,*, *59*, 381–384.
Barabasi, A. L. (2002). *Linked: The new science of networks*. Cambridge/Massachusetts: Perseus.
Condon, E. V. (1928). Statistics of vocabulary. *Science*, *67*, 300.

de Marchi, M., & Rocchi, M. (2001). The editorial policies of scientific journals: Testing an impact factor model. *Scientometrics*, *51*(2), 395–404.
Dewey, G. (1923). *Relative frequency of English speech sounds*. Cambridge/Massachusetts: Harvard University Press.
Dresden, A. (1922). A report on the scientific work of the Chicago section, 1897–1922. *Bulletin of the American Mathematical Society*, *28*, 303.
Egghe, L. (2005). *Power laws in the information production process: Lotkaian informetrics*. Amsterdam: Elsevier.
Egghe, L., & Rousseau, R. (1990). *Introduction de Informetrics*. Amsterdam: Elsevier.
Estoup, J. B. (1916). *Gammes stenographiques*. Paris: Institut Stenographique de France.
Fröhlich, G. (1996). The surplus value of scientific communication. *Review of Information Science*, *1*(2), 1–13.
Garfield, E. (1994). The impact factor. *Current Contents*, *25*, 3–7.
Langton, Ch. G. (1990). Computation at the edge of chaos. *Physica D*, *42*, 12–37.
Le Quan, H., Sicilia-García, E. I., Ming, J., & Smith, F. J. (2002). Extension of Zipf's law to words and phrases. In *Proceedings of the 17th International Conference on Computer Linguistics*.
Li, W. (1991). Expansion-modification systems: A model for spatial $1/f$ spectra. *Physical Review E*, *43*, 5240.
Li, W. (2003). http://linkage.rockefeller.edu/wli/zipf/.
Lotka, A. J. (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Science*, *16*, 317.
Mandelbrot, B. B. (1954). Structure formelle des textes et communication. *Word*, *10*, 1–27.
Mansilla, R., & Cocho, G. (2000). Multiscaling in expansion-modification systems: An explanation for the long-range correlation in DNA. *Complex Systems*, *12*, 207.
McElvey, B. (2001). What is complexity science? *Emergence*, *3*, 137–157.
Merton, R. K. (1968). The Matthew effect in science. *Science*, *159*, 56–63.
Nature. (2005). Editorial. *Nature, 435*, 1003–1004.
Popescu, I. (2003). On a Zipfs Law Extension to Impact Factors. *Glottometrics*, *6*, 83–93.
Simon, H. (1955). On a class of skew distribution functions. *Biometrika*, *42*, 425.
Simon, H. (1957). *Models of man*. New York: Wiley and Sons.
Willis, J. (1922). *Age and area*. Cambridge, UK: Cambridge University Press.
Yule, G. (1924). A mathematical theory of evolution. *Philosophical Transactions of the Royal Society*, *213*, 21–87.
Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Cambridge/Massachusetts: Addison-Wesley Press.

Available online at www.sciencedirect.com
ScienceDirect
Journal of Informetrics 1 (2007) 170–177
Journal of INFORMETRICS An International Journal
www.elsevier.com/locate/joi

On the *h*-index, the size of the Hirsch core and Jin's *A*-index

Quentin L. Burrell

Isle of Man International Business School, The Nunnery, Old Castletown Road, Douglas, Isle of Man IM2 1QB, via United Kingdom

Received 7 November 2006; received in revised form 4 January 2007; accepted 5 January 2007

Abstract

Hirsch's *h*-index seeks to give a single number that in some sense summarizes an author's research output and its impact. Essentially, the *h*-index seeks to identify the most productive core of an author's output in terms of most received citations. This most productive set we refer to as the Hirsch core, or *h*-core. Jin's *A*-index relates to the average impact, as measured by the average number of citations, of this "most productive" core. In this paper, we investigate both the total productivity of the Hirsch core – what we term the size of the *h*-core – and the *A*-index using a previously proposed stochastic model for the publication/citation process, emphasising the importance of the dynamic, or time-dependent, nature of these measures. We also look at the inter-relationships between these measures. Numerical investigations suggest that the *A*-index is a linear function of time and of the *h*-index, while the size of the Hirsch core has an approximate square-law relationship with time, and hence also with the *A*-index and the *h*-index.

Keywords: Hirsch *h*-index; Hirsch *h*-core; Jin *A*-index; Stochastic model; Informetric process

1. Introduction

Ever since Hirsch (2005) proposed the *h*-index, a single number to measure both an individual's research output and its impact, it has received much attention, both in the popular domain and in the academic literature; see Burrell (2007a) for some references. Jin (2006) has suggested a supplementary measure, termed the *A*-index – since it relates to an average – by Rousseau (2006). In this note, we investigate the *A*-index and what we term the Hirsch core using the model proposed by Burrell (1992, 2007a) for the publication–citation process. More recently, Egghe (2006) has proposed the *g*-index and Kosmulski (2006) the $h(2)$ index. Although, still seeking to identify a core of an author's published work based upon citation counts, these latter two are suggested as alternatives to the *h*-index and will be discussed elsewhere (Burrell, 2007b,c). All of these measures seek to provide simple summary statistics for an author's impact, in which case it is useful for the scientometrician to know how they might depend upon both the internal – author based – and external – environment based – influences.

2. The stochastic model

Here, we just recap the essentials of the model and refer the reader to Burrell (2007a) for full details. The basic idea is that an author publishes papers at certain times and that these papers subsequently attract citations following their publication, where both the publication and citation accumulation processes are random. We further assume that

E-mail address: q.burrell@ibs.ac.im.

1751-1577/$ – see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.joi.2007.01.003

some papers are more citable than others so that the citation rate varies between different publications. Much of this was originally described by Burrell (1992). The precise technical assumptions, without the mathematical details, are.

2.1. Assumptions

(1) From the start of his/her publishing career at time zero, an author publishes papers according to a Poisson process of rate θ, which gives the mean number of publications per unit time, called the *publication rate*.
(2) Any particular publication acquires citations according to a Poisson process of rate Λ, where Λ varies from paper to paper. Here, Λ denotes the mean number of citations to the paper per unit time following publication, called the *citation rate*.
(3) The citation rate Λ for this author varies over the set of his/her publications according to a gamma distribution of index $\nu \geq 1$ and scale parameter $\alpha > 0$.

See Burrell (2007a) for the precise details.

Remarks.

(a) Although, the citation rate depends on the two gamma parameters, α and ν, Burrell (2007a) found that his results were fairly robust to changes in the two parameters so long as the mean, i.e. the ratio $\mu = \nu/\alpha$ of the parameters, remains the same. Our numerical investigations for both the size of the h-core and the A-index are similarly fairly robust, i.e. the details change slightly but the general picture is the same. In all that follows we use, for purposes of illustration, citation rates of $\mu = 2$, 5 and 10, where the actual calculations are performed using $\alpha = 1$.
(b) As a referee has pointed out, one can argue over the robustness of the model assumptions – indeed, see Burrell (2007a) for some reservations – but at least they lead to a simple stochastic model that is analytically viable. In particular, the assumption of a fixed publication rate has been questioned by Burrell (2007b), based on Liang (2006) empirical data.

The basic result for the model is the following:

Theorem (Burrell, 2007a). Under the assumptions of the model, the distribution of X_T, the number of citations to a randomly chosen paper by time T, is given by

$$P(X_T = r) = \frac{\alpha}{(\nu - 1)T} B\left(\frac{T}{\alpha + T};\quad r + 1,\quad \nu - 1\right) \text{ for } r = 0, 1, 2, \ldots \tag{1}$$

where $B(x; a, b) = [\Gamma(a + b)/\Gamma(a)\Gamma(b)] \int_0^x y^{a-1}(1 - y)^{b-1}\, dy$ is the cumulative distribution function of a beta distribution (of the first kind) with parameters a and b.

3. Time-dependence of the size of the Hirsch core

According to the preprint of Hirsch (2005), the h-index for an author is that integer h such that h of his/her papers have at least h citations each, while the rest have fewer than h citations. Actually, this is not quite well-defined, see the print version of Hirsch (2005), Glänzel (2006) and Rousseau (2006), since there is ambiguity if there are several papers with the same number of citations at h. To get round this, let us introduce

Notation. Write $f(n; T)$ for the number of an author's papers receiving exactly n citations by time T, and $N(n; T)$ for the number of an author's papers that have received at least n citations by time T so that $N(n; T) = \sum_{r=n}^{\infty} f(r; T)$.

Definition 1. Hirsch's h-index at time T is, for any particular author, the integer $h(T)$ satisfying

$$h(T) = \max\{n : n \leq N(n; T)\}$$

Note that this is an empirical measure, requiring observation of the actual values of $N(n; T)$.

Rousseau (2006) has proposed the idea of the *Hirsch core* as the set consisting of the first $h(T)$ articles, in order of decreasing citations. In case there are more than one articles with $h(T)$ citations, he proposes listing them in anti-chronological order since this rewards the newer articles, which have achieved their $h(T)$ citations in a shorter time. Alternatively, one could argue that the Hirsch core should comprise all of the author's publications that have received at least $h(T)$ citations thus, perhaps, including some older papers. Note that his distinction, although possibly important in empirical studies, will not affect our theoretical development.

Definition 2. The *size* of the Hirsch core at time T is denoted by $C(T)$ and gives the total number of citations accumulated by those papers in the Hirsch core. Thus

$$C(T) = \sum_{n=h(T)}^{\infty} nf(n;T)$$

Note that this is again an empirical definition, requiring the observed numbers of citations $f(n;T)$ of those papers in the core.

Remarks. The h-index relates to the number of *publications* in the core, the size gives the total number of *citations* accumulated by the publications in the core. It has been pointed out by Hirsch (2005), Egghe (2006) and Rousseau (2006) that all one can say definitely is that this total is at least h^2, since there are h papers having at least h citations each. Thus, $C(T) \geq h(T)^2$. Within the limits of our model assumptions, we shall see that we are able to estimate the size, i.e. the (theoretical) number of citations to papers in the Hirsch core.

We calculate the theoretical, i.e. the expected value of, $C(T)$ by means of the following, which gets round the difficulty that the above definition involves an infinite sum:

Proposition. The expected total number of citations accumulated by those papers in the Hirsch core is, for our assumed model

$$E[C(T)] = \theta T \left(\frac{\nu T}{2\alpha} - \frac{\alpha}{(\nu-1)T} \sum_{n=0}^{h(T)-1} nB\left(\frac{T}{\alpha+T}; n+1, \nu-1 \right) \right) \tag{2}$$

Proof. See Appendix A.

Remarks.

(i) Although this expression may look unwieldy, its basic form is intuitively reasonable. An author producing on average θ publications per unit time over a period of length T will (on average) have produced a total of θT papers, each receiving (on average) ν/α citations per unit time. The average time for a paper to be available for citation is $T/2$ so that the total number of accumulated citations will be (on average) the product of these, namely $\theta\nu T^2/2\alpha$. This is the main term on the RHS of (2); the other is the deduction for total citations received by those papers not in the Hirsch core. (The precise mathematical justification of this intuitive argument is found in Appendix A.) Note that, already, this suggests that the size of the core could correlate, at least approximately, with the square of time.

(ii) The $h(T)$ in the above is now the theoretical h-index, determined as described in Burrell (2007a).

(iii) The actual calculation of the RHS of (2) is now straightforward, for any given set of parameter values, with any computer package allowing evaluation of the cumulative beta distribution.

(iv) At first sight, it might appear from (2) that $E[C(T)]$ is directly proportional to the publication rate θ. However, the range of the summation that appears on the RHS of (2) involves $h(T)$ and this in turn depends on θ. In fact, Burrell (2007a) conjectures that $h(T)$ is approximately linear in $\ln\theta$.

Although, it does not seem possible to give a straightforward analytical description of the time-dependence of the size of the h-core, the model does allow numerical investigation to at least suggest the form of the dependence. Purely for purposes of illustration, we have chosen to fix the publication parameter at $\theta = 5$, representing a "moderate" rate.

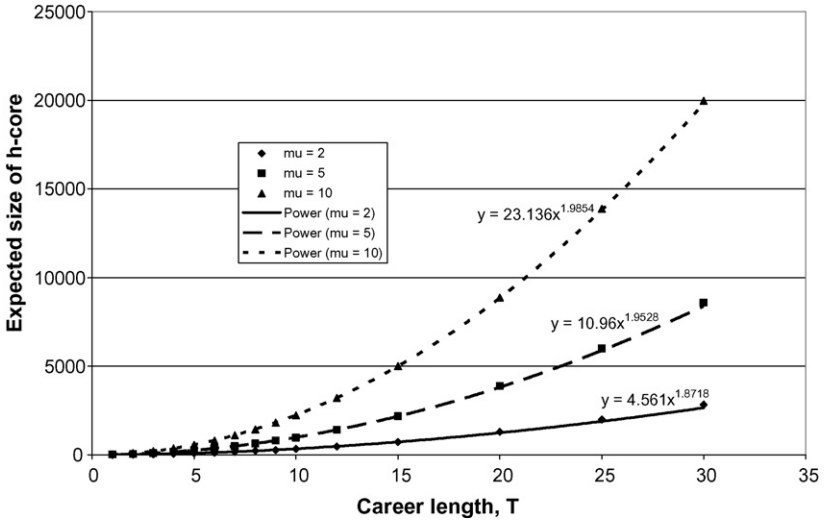


Fig. 1. Growth of h-core with time. Mean publication rate = 5.

Of course an average of five publications per year would be viewed as high in fields such as mathematics and possibly as low in other fields reporting a high level of collaborative work. For the citation rate, which can be thought of as the environmental or external factor, we use $\mu = 2$, 5 and 10 which correspond to low, medium and high rates of citation. The relationship between time and the size of the h-core for these scenarios is illustrated in Fig. 1, where the plotted points correspond to the time points or (current) career lengths $T = 1$–10, 12, 15, 20, 25 and 30.

Remarks. From Fig. 1, we can see quite clearly a very close power-law relationship between the size of the h-core and time. In all cases, the fit is extremely good, with $R^2 > 0.99$, and, in particular, note that the actual power is approximately (but always slightly less than) two. This last point should not be too surprising. We have already established that the expected *total* number of citations is proportional to T^2, it then seems reasonable that this should at least approximately be the case also when restricting attention to the core sources.

4. Time-dependence of Jin's A-index

According to Rousseau (2006), Jin's idea of an A-index (Jin, 2006) is that it should be the average number of citations received by those publications in an author's Hirsch core. In our notation, then, Jin's time-dependent A-index is as in:

Definition 3. Jin's A-index at time T is given by

$$A(T) = \frac{C(T)}{h(T)} = \frac{1}{h(T)} \sum_{n=h(T)}^{\infty} n f(n;T)$$

Again, this is an empirical measure, whereas we are investigating a theoretical model so we modify this to:

Definition 4. The theoretical A-index at time T is given by

$$A(T) = \frac{E[C(T)]}{h(T)} = \frac{1}{h(T)} \sum_{n=h(T)}^{\infty} nE[f(n;T)] = \frac{\theta T}{h(T)} \sum_{n=h(T)}^{\infty} nP(X_T = n)$$

where $h(T)$ is now the theoretical h-index, calculated as in Burrell (2007a).

As this again involves, the evaluation of an infinite sum, for purposes of calculation we use the following straightforward:

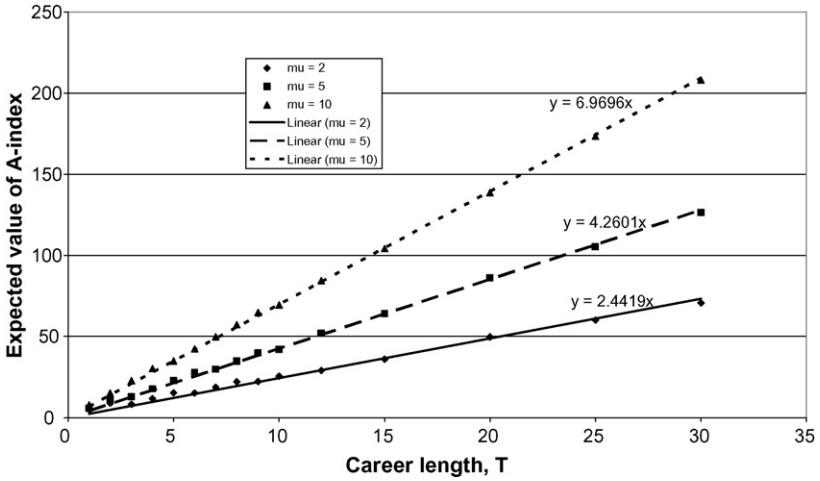


Fig. 2. Growth of A-index with time. Mean publication rate = 5.

Corollary to the Proposition. Under the assumptions of the model, the theoretical A-index is given by substituting the expression for $E[C(T)]$ from (2) into the basic definition of $A(T)$ above.

$$A(T) = \frac{\theta T}{h(T)} \left(\frac{\nu T}{2\alpha} - \frac{\alpha}{(\nu - 1)T} \sum_{n=0}^{h(T)-1} nB\left(\frac{T}{\alpha + T}; n + 1, \nu - 1 \right) \right)$$

Although, again, it does not seem possible to give a direct analytic expression of the dependence between the A-index and time, evaluation of the above expression for any given parameter values is routine given a computer package including the cumulative beta distribution function. Using the same combinations of parameter values as before, we illustrate the results of such calculations in Fig. 2.

Note that, as it seems intuitively reasonable that at time $T = 0$ the A-index should also be equal to zero, the fitted line – and the displayed regression equation – has the constraint that it should pass through the origin. In all cases, the approximate linearity is evident; indeed, we again have $R^2 > 0.99$. The reason for the linearity can be explained as follows. By its definition, Jin's A-index is given by $A(T) = C(T)/h(T)$. But we have just argued that $C(T)$ is (approximately) proportional to T^2 and Burrell (2007a) investigations strongly suggest that, with this model, $h(T)$ is approximately proportional to T. Taking these together, we should expect that A, as their ratio, is also approximately proportional to T.

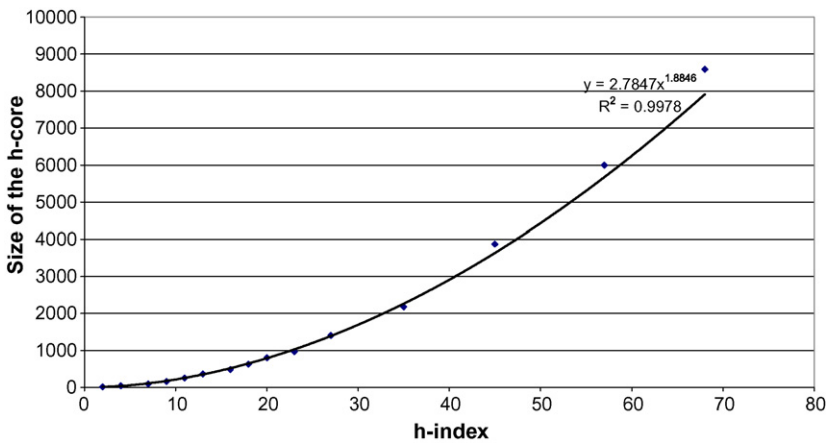


Fig. 3. Growth over time of the h-core with the h-index.

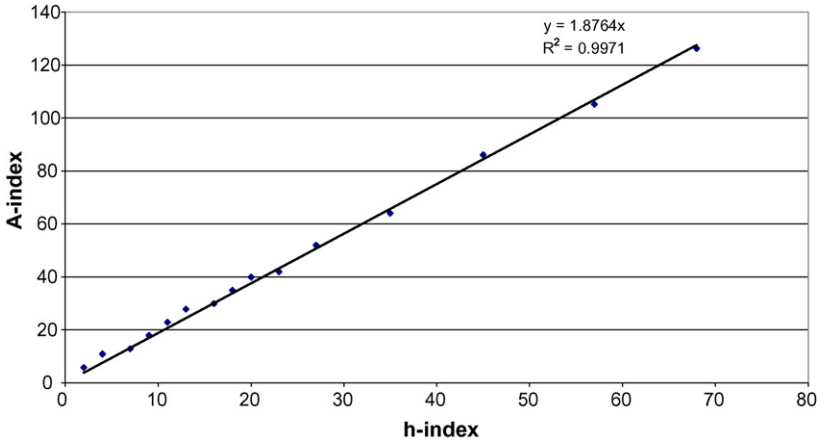


Fig. 4. Growth over time of A-index with h-index.

5. Relationships with the h-index

Given that $C(T)$ is approximately proportional to T^2 and that $h(T)$ is approximately proportional to T (Burrell, 2007a), it is not hard to see that, according to this model, the size of the core should be, at least approximately, proportional to h^2. Similarly, since $A(T)$ is approximately proportional to T, we would expect that the A-index is approximately proportional to the h-index. We illustrate these relationships in Figs. 3 and 4, which confirm these arguments. (Note that we have used a publication rate of $\theta = 5$ and citation rate $\mu = 5$ for these graphs. Other values produce similar results, but with different constants of proportionality.)

In Fig. 3, the reader might argue that, although the reported value of R^2 is very high, visual inspection suggests that the divergence increases with increasing h. In fact this results from our original (restricted and unequal) choice of values for the time parameter/career length T.

6. Concluding remarks

We have shown that Burrell's (1992, 2007a) model for the publication–citation process allows analytic and numerical investigation of the time-dependent behaviour of both the size of the h-core and the A-index (for an individual author). The main results are that the model suggests that the size of an author's Hirsch core should be approximately proportional to the square of the h-index (and of time), while for the A-index we should expect approximate direct proportionality to h and time. We await empirical studies to see to what extent these general findings agree with practice. Such studies would be analogous to Liang (2006) work on the time evolution of an author's h-index, but working forwards from the beginning of an author's active career, not backwards, see Burrell (2007b) for comments on this. Our own feeling is that all of these measures are indicative, supplementary scientometric measures. It will take much more empirical as well as theoretical research before anyone can claim a single definitive measure.

Appendix A

Proof of proposition. From its definition, we have

$$E[C(T)] = E\left[\sum_{n=h(T)}^{\infty} n f(n;T)\right] = \sum_{n=h(T)}^{\infty} n E[f(n;T)] = \theta T \sum_{n=h(T)}^{\infty} n P(X_T = n) \tag{A1}$$

$$\text{But} \quad \sum_{n=h(T)}^{\infty} n P(X_T = n) = E[X_T] - \sum_{n=1}^{h(T)-1} n P(X_T = n) \tag{A2}$$

So far as $P(X_T = n)$ is concerned, this follows from Eq. (1). Hence, all we need to complete the proof is an expression for the mean of X_T. This is given by the following:

Lemma.

$$E[X_T] = \frac{\nu T}{2\alpha}$$

(i) Standard proof

By definition

$$E[X_T] = \sum_{r=0}^{\infty} rP(X_T = r) = \sum_{r=1}^{\infty} rP(X_T = r) = \frac{\alpha}{(\nu - 1)T} \sum_{r=1}^{\infty} rB\left(\frac{T}{\alpha + T}; r + 1, \nu - 1\right)$$

making use of Eq. (1).

For the summation, making use of the integral representation of the cdf of the beta distribution, we have

$$\begin{aligned}\sum_{r=1}^{\infty} rB\left(\frac{T}{\alpha + T}; r + 1, \nu - 1\right) &= \sum_{r=1}^{\infty} r\left(\frac{\Gamma(r + \nu)}{\Gamma(r + 1)\Gamma(\nu - 1)} \int_0^{T/(\alpha+T)} y^r (1 - y)^{\nu-2}\, \mathrm{d}y\right) \\ &= \int_0^{T/(\alpha+T)} \left(\sum_{n=0}^{\infty} \frac{\Gamma(n + 1 + \nu)}{n!\Gamma(\nu - 1)} y^{n+1}(1 - y)^{\nu-2}\right) \mathrm{d}y, \quad \text{where} \quad n = r - 1 \\ &= \nu(\nu - 1) \int_0^{T/(\alpha+T)} \frac{y}{(1 - y)^3} \left(\sum_{n=0}^{\infty} \frac{\Gamma(n + (\nu + 1))}{n!\Gamma(\nu + 1)} y^n (1 - y)^{\nu+1}\right) \mathrm{d}y\end{aligned}$$

Now recognise the inner summation as the total sum of the probability mass function of NBD$(1 - y, \nu + 1)$ random variable and hence is equal to 1 for any y in [0,1].

It is then routine calculus to show that

$$\int_0^{T/(\alpha+T)} \frac{y}{(1 - y)^3}\, \mathrm{d}y = \frac{T^2}{2\alpha^2}$$

Substituting back, then, we find that

$$E[X_T] = \frac{\alpha}{(\nu - 1)T} \nu(\nu - 1) \frac{T^2}{2\alpha^2} = \frac{\nu T}{2\alpha} = \frac{\nu}{\alpha} \frac{T}{2}$$

(ii) Smart proof

$E[X_T] = E_t E[X_T|t]$, where t denotes the (random) time at which the typical paper was published. Now, given the publication time t, the paper has been in the public domain for a time $T - t$ gathering citations at expected rate ν/α per unit time so that $E[X_T|t] = (T - t)\ \nu/\alpha$

$$\text{Thus,} \quad E[X_T] = E_t E[X_T|t] = E[(T - t)\nu/\alpha] = (T - E[t])\nu/\alpha$$

But according to the model, publications appear as a Poisson process and hence publication times are uniformly distributed over [0,T], see for instance, Ross (1996, Chapter 2, p. 67) or Stirzaker (2005, Chapter 2, p. 75). In particular, then, $E[t] = T/2$ and the result follows.

It is now a straightforward matter to substitute this into (A2) and hence (A1). Making use of (1) we establish the result as in (2).

References

Burrell, Q. L. (1992). A simple model for linked informetric processes. *Information Processing and Management*, *28*, 637–645.

Burrell, Q. L. (2007a). Hirsch's h-index: A stochastic model. *Journal of Informetrics*, *1*(1), 16–25.

Burrell, Q. L. (2007b). Hirsch index or Hirsch rate? Some thoughts arising from Liang's data. *Scientometrics*, *73*(1), submitted for publication.

Burrell, Q. L. (2007c). Hirsch's h-index and Egghe's g-index. To be presented at the 11th ISSI Conference, 25–27 June, Madrid, Spain.

Egghe, L. (2006). Theory and practice of the *g*-index. *Scientometrics*, *69*(1), 131–152.

Glänzel, W. (2006). On the *h*-index—A mathematical approach to a new measure of publication activity and citation impact. *Scientometrics*, *67*(2), 315–321.

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*(46), 16569–16572. (Also available in preprint form as ar Xiv: physica/0508113, accessible at http://xxx.arxiv.org/abs/physics/0508025).

Jin, B. H. (2006). *h*-Index: An evaluation indicator proposed by scientist. *Science Focus*, *1*(1), 8–9. (In Chinese)

Kosmulski, M. (2006). A new Hirsch-type index saves time and works equally well as the original *h*-index. *ISSI Newsletter*, *2*(3), 4–6.

Liang, L. (2006). *h*-Index sequence and *h*-index matrix: Constructions and applications. *Scientometrics*, *69*(1), 153–159.

Ross, S. (1996). *Stochastic processes* (2nd ed.). New York: John Wiley.

Rousseau, R. (2006). New developments related to the Hirsch index. *Science Focus*, *1*(4), 23–25. (In Chinese).

Stirzaker, D. (2005). *Stochastic processes and models*. Oxford: Oxford University Press.