Jointly published by Elsevier Science Ltd, Oxford and Akadémiai Kiadó, Budapest

Scientometrics, Vol. 35, No. 3 (1996) 375-386

RELATIONSHIPS BETWEEN THE RATE OF SCIENTIFIC DEVELOPMENT AND CITATIONS. THE CHANCE FOR CITEDNESS MODEL

P. VINKLER

Central Research Institute for Chemistry, Hungarian Academy of Sciences, Pusztaszeri út 59-67, 1025 Budapest (Hungary)

(Received Oktober 5, 1995)

Chances for information to be cited (CC) depend on disciplines and topics because of different publication and referencing practices. However, the developmental rate of knowledge strongly influences CC as well. By a simple model it has been concluded that CC are the greater the faster the publication rate.

Introduction

Citations and references represent basic categories in Scientometrics. Citations received by papers may indicate the impact of results published whereas references given may characterize the information base of referencing authors.[1]

Citation and publication lists and indicators derived from them play an increasing role in obtaining research grants, positions, fellowships etc. Obtaining more citations has become a crucial point for every scientist and research organisation. Investigations on motivations of referencing revealed that scientific relevancy and quality of the published results may be responsible for about 70 per cent of references whereas 30 per cent may be attributed to indirect scientific and non-scientific reasons.[2] The *chance to be cited* is enhanced *a priori* by increased *relevancy* and higher *quality* of the potentially citable publications, but quantification of the mentioned *inherent* characteristics is difficult.

It is well known that there are great differences concerning publishing and referencing habits and traditions by fields, subfields or topics.[3] From the data of *impact factors* of journals (i.e. mean citedness of papers in a given journal) representing different subfields it was concluded[4] that the main factors determining the citedness of journals are as follows:

0138-9130/96/US $ 15.00

- mean number of references by papers,
- extent of interdisciplinary connections,
- ratio of newer and older references.

The *distribution of references by age* strongly depends on the *rate of development* of topics, subfields or fields.[5] The high value of the Price index (ratio of references published in the most recent five year period to the total) of a subfield may indicate high growth rate.

It was found earlier[4] that the measure of disciplines does not influence the chances to be cited calculated for a given state. Therefore, it seemed to be interesting to investigate the change in citedness possibilities in time.

The growth rate of the scientific literature in time (t) can be described by the increase in the number of scientific publications (N) (Eq. 1).

$$(dN/dt) = kN \qquad \text{(Eq. 1)}$$

where k is a constant.[6] The equation shows that the increase of information depends on the starting level of the information pool (N). The solution of the differential equation with the initial condition of $N = N_0$ at time $t = 0$ is as follows:

$$N(t) = N_0 e^{kt} \qquad \text{(Eq. 2)}$$

The exponential curve obtained by Eq. (2) can be characterized by the time during which N is doubled. The growth rate, however, may be exponential only until external conditions of the information system do not change. Limiting factors make the function of increase logistic. The logistic developmental curves of a given system can build up to each other investigating longer time-periods.

The increase in the knowledge of scientific disciplines could be demonstrated by the numbers of scientific papers issued yearly. Braun, Bujdosó and *Schubert*[7] offer several examples showing the different increase rates of different scientific fields or topics in time. The doubling time (2T) of the quantity of information was found to be very different (1.5 years $\leq$ 2T $\leq$ 25 years) depending on topic and *time*. The yearly production of scientific papers ranges from several tens up to several thousands or ten thousands, strongly depending on the extent and features of the field selected and the *time-period* of the investigation. The growth of the number of papers in all branches of chemistry all over the world shows a 2T value of 14.5 years between 1910-1970.[7]

Price[8] gives several examples that "...science increases in all its aspects exponentially." Some recent examples, however, show that the increase reached saturation state on several fields and in several aspects.

The annual change in the total R&D spending (defense included), for instance, was found to be 3, 6, 4, 3, 2, 7 and 4 per cent for U.S., Japan, Germany, France, U.K., Italy and Canada, respectively between 1982-1992.[9]

The annual change in the federal support for scientific disciplines in U.S. took 7 per cent between 1987-1994. Metallurgy and Materials led by 13 per cent, Environmental Sciences followed by 8 per cent whereas Chemistry, Life Sciences and Physics took 6 per cent.[9]

The annual change in the number of graduate science students was only 2 per cent in the same period. The highest change was found for Metallurgical and Materials Engineering (6), whereas Chemistry, Biochemistry, Physics and Mathematical and Computer Sciences took 2, 3, 3 and 4 per cent, respectively.

Results and discussion

The increase in the scientific information by disciplines or topics may be represented by *information packages* containing the total of scientific publications issued worldwide, yearly. The number of publications can be approximated by that of scientific papers in journals. The numbers of papers published in each year may represent *special mathematical series*.

In the following it is attempted to introduce a simple *model of changing in the citedness probabilities of papers depending on the rate of change in information quantity* produced from year to year.

Citedness, which is a specific measure, means here a ratio, namely, the number of citations obtained in a selected time-period by a single paper or a set of papers published in a given time-period divided by the number of those papers. Obtaining a citation by a paper corresponds to producing a reference by another paper at the same time. Consequently, citations may indicate the impact of the published information.

It is well known that Garfield impact factors of journals[10] depend strongly on fields and subfields. It was attempted to introduce "Subfield Factors" in order to normalize the different mean impact factors of journals on different subfields.[3]

In order to characterize the probability for obtaining citations in a single year on fields with different rates of development the *Chance for Citedness* index, CC(t) has been introduced (Eq. 3).

$$CC(t) = [N_r(t+1)r]/N_p(t) \qquad \text{(Eq. 3)}$$

where N_p is the total number of papers published earlier (potentially "citable items") in the publication period (t) selected and N_r is the number of the potentially referencing papers published in a consequent single year $(t+1)$, whereas r is the mean number of references in the respective papers. (N_r multiplied by r gives the total number of references whereas N_p is the number of papers to be referenced.)

The total number of citations to be received during the life time (T) of papers (*Total Citedness Possibility*, TCP (T) can be calculated as follows:

$$TCP(T) = CC(t) \times T \qquad \text{(Eq. 4)}$$

where CC(t) is the respective Chance for Citedness index and T is the mean lifetime (in years) of information on the respective field or subfield. *A priori*, the publication time-period (t) (i.e. production period of "citable items") can be shorter or longer than or equal to the life-time (T) of information in papers.

In order to arrive at any conclusion concerning the application of CC(t) indicators calculated for fields with different developmental rates, some model experiments should be carried out. Starting assumptions to the model are as follows.

Basic assumptions to the Chance for Citedness Model

1.) The information in papers published in each year (N_k) is supposed to be relevant (valid) for a period of *two, five or ten years*, respectively (i.e. the life-time of the papers (T) is equal to two, five, or ten years).
2.) Papers (N_r) published in a selected year ($t+1$; termed as *referencing year*) exclusively reference papers (N_p) published in the preceding two, five or ten years ($t=2$, 5, 10 years) period, respectively.
3.) All papers reference only those published on the same subfield.
4.) All referencing papers (N_r) reference a single paper published earlier (i.e. $r_i=r=1$).

Table 1

Chance for Citedness CC(t) indices for scientific fields (A, B, C, D, E) with different publication rates

year	Number of papers published yearly (N_k)				
(k)	A	B	C	D	E
1.	100	100	100	100	100
2.	100	110	200	400	90
3.	100	121	300	900	80
4.	100	133	400	1600	70
5.	100	146	500	2500	60
6.	100	161	600	3600	50
7.	100	177	700	4900	40
8.	100	195	800	6400	30
9.	100	215	900	8100	20
10.	100	237	1000	10000	10
N_r (3)	100	121	300	900	80
N_r (6)	100	161	600	3600	50
N_r (11)	100	261	1100	12100	10
N_p (2)	200	210	300	500	190
N_p (5)	500	610	1500	5500	400
N_p (10)	1000	1595	5500	38500	550
CC (2)	0.500	0.576	1.000	1.800	0.421
CC (5)	0.200	0.263	0.400	0.654	0.125
CC (10)	0.100	0.163	0.200	0.314	0.018

Legends:

A: $N_k = 10^2$

B: $N_{k+1} = N_k + \frac{N_k}{10}$

C: $N_k = 10^2 k$

D: $N_k = (10k)^2$

E: $N_{k+1} = N_1 - 10k$

N_r (t+1): number of referencing papers published in a year (i.e. in years 3, 6 and 11, respectively)

$N_p(t)$: total number of papers published during period t (i.e. during 2, 5 and 10 years, respectively).

$$CC(t) = \frac{N_r(t+1)}{N_p(t)}$$

Ad 1 and 2

The calculation of *impact factors* as suggested by *Garfield*[10] applies *two year* publication and a consequent single year citation period. The *Price index*[5] gives the share of the papers referenced which were published during a *five year* period prior to the publishing year. The time period recorded in *Science Citation Index, Journal Citation Reports, Citing Journal Package*[11] spans ten years. The facts mentioned indicate to select periods of 2, 5 and 10 years for investigating citedness possibilities. The Price index is believed to characterize the research front in terms of immediacy of references, whereas impact factors of journals are widely used for investigating journal characteristics including recency, eminency, utility etc. Stability of information over time can be traced by studying longer time intervals (5 or 10 years).

Table 2

Total Citedness Possibility, TCP(T) and Standardized Chance for Citedness, SCC(t) values calculated with life-times, T=2, 5 and 10 years, respectively, for different scientific fields (A, B, C, D, E)

	A	B	C	D	E
TCP(2)	25	28.8	50	90.0	21.0
TCP(5)	25	33.0	50	81.8	15.6
TCP(10)	25	40.9	50	78.6	4.5
SCC(2)	1.00	0.87	0.50	0.28	1.19
SCC(5)	1.00	0.76	0.50	0.31	1.60
SCC(10)	1.00	0.61	0.50	0.32	5.56

Legends:

For A, B, C, D, E see Table 1

TCP (T) values are calculated with r=25 for all fields.

$$TCP(T) = \frac{25 \cdot N_r(t+1)}{N_p(t)} T \; ; \qquad SCC(t)_A = \frac{CC(t)_A}{CC(t)_x}$$

For $N_r(t+1)$ and $N_p(t)$ values see Table 1

x is for A, B, C, D, E, respectively.

Ad 3

Information relations between scientific subfields are very different. The assumption is based on the fact *that papers in a journal preferably reference publications appeared in the same journal.*[3] This observation is preferably valid for journals ranked high in impact factor lists. Consequently, it is reasonable to assume that the majority of information is applied by the researchers working on the same subfield as the information producers.

Ad 4

As first approximation it was assumed that a referencing paper could reference only a single earlier paper. Total Citedness Possibility (TCP) values taking into account citations obtained during the life-time of papers are given later (see Table 2).

Five publishing rates were selected for demonstrating the development of research fields (Table 1). *A* is a field with constant rate (dN/dt = constant) producing 100 papers in each year (N_k). *B* shows 10 per cent increase from year to year, the increase in the number of papers produced annually is 100 for field *C*, whereas *D* represents a quadratic increase. Field *E* shows a decrease in the annual number of papers.

$N_r(t+1)$, $N_p(t)$ and the respective CC(t) values are given in Table 1, from which the following conclusions can be drawn.

Characteristics of the Chance for Citedness, CC(t) indices

1.) The CC(t) indices are greater for fields with greater publication rate (D>C>B>A>E).
2.) The CC(t) indices are the smaller the greater the life-times (T) of the papers: CC(2)>CC(5)>CC(10).
3.) The measure of CC(t) indices within a field can be independent on the selected referencing and publication time-periods only if Eq. (5) holds.

$$N_{k+1} / \sum_{i=1}^{k} N_i = N_r(t+1)/N_p(t) = \text{constant} \qquad \text{(Eq. 5)}$$

where N_i is the number of papers in the i-th year, $N_r(t+1)$ is the number of referencing papers and $N_p(t)$ is the number of papers to be referenced.

Some practical examples

The main conclusion of the Chance for Citedness (CC) Model is that *chances to be cited would be increased on fields with increasing publication rate*. Motivations of individual researchers for making references, reasons for referencing or neglecting individual publications are far from complete understanding yet. The citedness of papers on a given subfield may be determined by several factors (e.g. relevancy of the information published; rate of the change in the number of publications in time; aging of information published; mean number of references per paper; multi- and/or interdisciplinarity of the information published; grade and type of the knowledge published.) The *separation of the effect of the individual factors* seems to be very difficult at present. Therefore, the model experiment presented was performed.

The verification of the CC Model described through citedness data of papers or journals needs complex investigations. Only some preliminary results can be given here.

Taking into account the change in the number of papers in time, three cases can be considered. The number of papers published on a subfield is yearly

- increasing (Table 1; B, C, D);
- constant (Table 1; A);
- decreasing (Table 1; E).

Another important factor influencing citedness of papers is the *aging* of the information published. This factor can be characterized by the *citing half-life* values of journals published in *Journal Citation Reports*.[11] (Citing half-life is the number of journal publication years from the current year going back which account for 50 per cent of the total references given by the citing journal in the current year.)

Price indices which give the ratio of the more recent 50 per cent of references are believed to represent *recency* of researches.[5] Citing half-life values may represent the same *expressed in time*. Consequently, the shorter is the citing half-life of a journal the greater the aging of the respective information and the higher the growth rate of the respective topic.

Our preliminary results show that journals on subfields developing faster (i.e. offering *shorter citing half-life periods*) have relatively *great impact factor* (i.e. greater chance to be cited) contrary to that with *smaller impact factor and longer life-time* data. Some examples (mean impact factors of journals dedicated to the respective

subfield) are given as follows: Chemistry, Physical: 1.78; Immunology: 2.55; Genetics and Heredity: 2.49; Physics, Atomic, Molecular and Chemical: 2.19; Oncology: 1.75. (The mean impact factors presented are calculated from data given in Ref. 11).

In contrast, there are subfields with journals of relatively *low impact factor* (i.e. lower chance for citedness) and *longer half-life periods*, e.g. Chemistry, Applied: 0.60; Geology: 0.96; History and Philosophy of Science: 0.35; Mathematics: 0.37; Paleontology: 0.64. Some fields with extremely great speed, however, show relatively low mean impact factor as well (e.g. Environmental Science: 0.70, Allergy: 0.89). This may be the consequence of a relatively great number of *new* journals on the field of which impact factor starts to increase only later.

Citing half-life values for the fields mentioned above are as follows: Chemistry, Physical: 6.60; Immunology: 4.78; Genetics and Heredity: 6.07; Physics, Atomic, Molecular and Chemical: 6.71; Oncology: 4.81; Chemistry, Applied: 7.10; Geology: 7.69; History and Philosophy of Science: 9.64; Mathematics: 9.07; Paleontology: 8.28.

Maximum possible number of citations

The maximum possible number of citations, TCP(T), which is available during the life-time (T) of scientific papers, depends on inherent (relevancy, quality, clearness etc.) and external factors. One of the latter mentioned factors is the mean number of references in the referencing papers.[4]

The Total Citedness Possibility (TCP) indicators calculated for fields in Table 1 are given in Table 2. Calculating with a mean number of references $r=25$, one can obtain 25 for any TCP(T) indicator for information systems of type A where the yearly production of publications (N_k) is *constant*. This is in accordance with the statement of *Plomp*[12] who concluded that *the total number of citations* received by an average paper over its life-time should be *equal to the mean number of references* (r) in referencing papers. This number, r was found to be 16 for Mathematics and Information Theory, 21 for Clinical Medicine, 25 for Physics and 28 for Chemistry and Biology.[12]

The TCP(T) values in Table 2, however, reveal that the *maximum number of citations* obtained by papers *can be equal to the mean number of references (r) only for papers published on fields with constant annual publication production* (like field *A*). Note that TCP(T) indices do not depend on the number of papers (N_k) produced annually at standard increase (e.g. C). In contrast, fields *B* and *D* show different numbers (Table 2). *A single TCP(T) index generally valid for the whole period cannot*

be calculated for systems like *D* or *E* because the value of the indicator depends on the time period selected. Some TCP(2) indicators for field *D* are e.g. as follows: $9000 \times 25/(100+400) = 45.0$; $1600 \times 25/(400+900) = 30.76$; $2500 \times 25/(900+1600) = 25.00$; $3600 \times 25/(1600+2500) = 21.95$ etc.

Table 3

Doubling times (2T) of information (in years) by fields (A, B, C, D) with different publication rates for different time periods

	A	B	C	D
$N_p(5)$	500	610	1500	5500
$2N_p(5)$	1000	1220	3000	11000
$N_p(6)$	600	771	2100	9100
$N_p(7)$	700	948	2800	14000
$N_p(8)$	800	1143	3600	-
$N_p(9)$	900	1358	-	-
$N_p(10)$	1000	-	-	-
2T(5)	5	3-4	2-3	1-2
$N_p(2)$	200	210	300	500
$2N_p(2)$	400	420	600	1000
$N_p(3)$	300	331	600	1400
$N_p(4)$	400	464	-	-
2T(2)	2	1-2	1	0-1

Legends:

$N_p(t)$: total number of papers published during period *t*
2T(t): doubling time of the number of papers in years, calculated from data referring to the respective *t* years period (i.e. 2 and 5 years, respectively)

In these cases special TCP(T) indices valid only for given years and time-periods should be calculated.

From the data of Table 1 it follows that chances of being cited are different for fields with different developmental rate. Therefore, the introduction of a *Standardized Chance for Citedness* (SCC) index is relevant within a system, for comparing different fields. It can be calculated as follows:

$$SCC(t)_A = CC(t)_A / CC(t)_X \qquad \text{(Eq. 6)}$$

where $CC(t)_A$ is the Chance for Citedness index obtained for the field with standard information production chosen as reference, whereas $CC(t)_X$ is that for any other field. The SCC(t) indices used as *multiplicative factors* can eliminate the discrepancies in citedness possibilities between fields or subfields with different rate of growth in information.

The SCC(t) indices for the fields presented in Table 1 are given in Table 2.

The rate of publication production is often characterized by the time (2T) needed for doubling the quantity of information on a subfield.[7] However, 2T values *depend on the time-period selected.* For doubling the information on field A, e.g. two or five years are needed considering the doubling of papers produced in the first two or in the first five years, respectively (Table 3). The data in Table 3 reveal that the *2T values decrease with increasing publication rate and increase with considering greater time-periods.*

Conclusions

Scientists working on different fields, subfields and topics are forced to publish increasingly more in order to obtain grants and positions. Selection of research topics depends on talent, knowledge and chance. Majority of the research workers, first of all younger people, however, are not in the position to choose fields of activity at their own discretion. It would be extremely important for every scientist to know the publication rate of the topic or subfield he or she is working on. The *publication rate* of subfields characterizes, namely, their *developmental phase*, which is one of the determining factors in *the number of citations attainable.*

The results of the present paper reveal that the Total Citedness Possibility TCP(T) values should be determined separately for each scientific field with a special publication rate and mean number of references. *Only scientific fields or subfields with the same rate of information production offer similar citedness possibilities.*

Consequently, only disciplines, fields or subfields approximately with the same mean number of references and similar development rates can be compared.

References

1. P. VINKLER, Literature overlap measures for information pools of research teams, In: *Proceedings of Fifth International Conference of the International Society for Scientometrics and Informetrics, Proceedings-1995*, M.E.D. KOENIG, A. BOOKSTEIN (Eds), Learned Information Inc., Medford NJ, 1995.
2. P. VINKLER, A quasi-quantitative citation model, *Scientometrics*, 12 (1987) 47.
3. P. VINKLER, Bibliometric features of some scientific subfields and the scientometric consequences thereform, *Scientometrics*, 14 (1988) 453.
4. P. VINKLER, Possible causes of differences in information impact of journals from different subfields, *Scientometrics*, 20 (1991) 145.
5. D. DE SOLLA PRICE, Citation measures of hard science, soft science, technology and non-science, In: *Communication Among Scientists and Engineers*, C.E. NELSON, D.K. POLLOCK (Eds), Heath Lexington, Mass., 1970.
6. V.V. NALIMOV, G.M. MULCHENKO, *Naukometriya*, (in Russian) Izd. Nauka, Moscow, 1969.
7. T. BRAUN, E. BUJDOSÓ, A. SCHUBERT, *Literature of Analytical Chemistry: A Scientometric Evaluation*, CRC Press, Boca Raton, 1987.
8. D. DE SOLLA PRICE, *Little Science, Big Science*, Columbia University Press, New York, 1961.
9. M.B. BRENNAN, J.R. LONG, *Facts & Figures for Chemical R&D, C&EN*, August 22, 1994.
10. E. GARFIELD, *Citation Indexing. Its Theory and Application in Science, Technology, and Humanities*, Wiley, New York, 1979.
11. *Science Citation Index, Journal Citation Reports, Citing Journal Package*, Institute for Scientific Information, Philadelphia, 1992.
12. R. PLOMP, The highly cited papers of professors as an indicator of a research group's scientific performance, *Scientometrics*, 29 (1994) 377.