Percentile Rank and Author Superiority Indexes for Evaluating Individual Journal Articles and the Author's Overall Citation Performance

A.I.Pudovkin

Institute of Marine Biology, Vladivostok 690041, Russia

aipud@mail.ru

E. Garfield

ThomsonReuters, Philadelphia, USA

garfield@codex.cis.upenn.edu

http://eugenegarfield.org

Presented at

Fifth International Conference on Webometrics, Informetrics & Scientometrics (WIS)

Tenth COLLNET Meeting, September 13-16, 2009 – Dalian, China

Abstract

In this paper we propose two new indexes to quantify the citation status of papers and authors. The Percentile Rank Index (PRI) indicates the citation rank of the author's individual papers among the papers published in the same year and source (journal or multi-authored monograph or book.) PRI is independent of the paper's age, specialty, or source journal size. The Author's Superiority Index (ASI) is determined by the number of the author's papers with a PRI at or above a specified value (99, 95, or 75). ASI allows comparisons across specialties and different time periods. The data necessary to calculate both the PRI and ASI can be obtained from Thomson-Reuters database Web of Science (www.isiknowledge.com) or other comparable databases.

The quantitative assessment of individual scientist's contributions is still a hot question. The now classic paper by Jorge Hirsch (2005) which proposed his h-index has already been cited in over 250 papers as of March 2009, and this number continues to grow. The advantages and drawbacks of the h-index are widely discussed. Many modifications of the h-index have been suggested. For a recent review see Bornmann & Daniel (2009).

To illustrate the problems with the h-index we provide citation data for two population geneticists (Table 1). They both have identical h-indexes (equal to 8), but quite different total citation numbers (87 and 391). From these data it is evident, that the h-index is not definitive or discriminating enough. Though Hirsch (2007) showed that the h-index is informative for predicting the future performance of scientists, the correlation of the values of h-index for two consecutive time periods for the same author is not strong. When examining a large group of scientists the h-index performs relatively well, but for individual authors it is not sufficiently predictive. However, it is probably impossible to characterize the impact of a scientist's contribution to science with a single number. Indeed, Hirsch states this explicitly in a personal communication (2008) with one of the authors.

All the citation indexes used for evaluatory purposes (there are many of them) are derivatives of the "raw" citation number. The latter is the number of papers, which cite the paper under consideration. Dealing with citation numbers we must remember that citedness of a paper

depends on many factors not related to its scientific quality or importance. Among these factors are the following: the age of the cited paper (which is the number of years after its publication), its specialty (science field), visibility, availability and prestige of the journal in which it was published. Though the latter factor seems to be partially correlated with paper quality: a manuscript of poor quality would most probably be rejected by a journal of high standing, that is having high impact factor (IF).

Table 2 illustrates dependence of paper citation rates on the paper age and its specialty. These data are obtained from Thomson's Essential Science Indicators database. One may see, that an average paper published in 1999 has been cited (by August, 2009) almost 18 times more than a paper, published in 2008 (A). An average paper in mathematics is cited 8 times less frequent than an average paper in molecular biology or genetics (B).

Table 3 confirms the fact of different citation rates in different science fields by considering journal IF in three fields. Again, mathematics is less cited field: the median value of IF of mathematics journals is almost 5 times less than the median IF of journals in Biochemistry and Molecular Biology.

Evidently, comparing citation performance of different authors we should take into consideration the ages of their papers, their specialty and visibility of journals, in which the papers are published. Thus, it seems highly desirable to work out a set of easily obtainable informative indicators, which are clear and transparent in meaning. They should be age, journal and specialy independent (or at least more independent than other similar indexes). We suggest such an indicator, the Author Superiority Index to complement the h-index and other citation indicators. The data necessary to calculate the index can be obtained from Thomson-Reuters database Web of Science (www.isiknowledge.com).

We propose a procedure involving a two step process. It requires that we first obtain a Percentile Rank Index (PRI) for each of the individual papers an author has published, and then calculate the Author Superiority Index (ASI), which is based on PRI values for all the author's papers. The PRI for each paper is based on the citation rank of the paper among the papers published in the same journal in the same year. In other words, the comparison is made among the related papers of the target one published within the same specialty journal or any topical group of papers one may aggregate by various methods, provided the papers are of the same age as the other papers under comparison. Thus, PRI also may be applied to papers published in multi-authored books, proceedings volumes, or other topical collections of papers. This suggestion is in line with our approach to characterizing journal impact factors (Pudovkin & Garfield, 2004). It is also relevant to mention that a similar journal-related paper citation ranking approach was used in the identification of "Citation Classics" published in Current Contents over a thirty year period (see the post: Garfield, 2008).

To illustrate how we obtain PRI values let us consider the data on h-core papers of an author in population genetics (Table 4). The h-core refers to the 29 papers that were cited 29 or more times, used to obtain the h-index. In table 4 the papers are sorted by citation frequency. The data were extracted from the Web of Science (WOS) database All the citation values dealt with in this paper refer to January 9, 2009. To retrieve the necessary data an author search is conducted to find all the papers of the specified author covered by WOS . For each paper one makes a journal search for a specific year and retrieves all the papers published by that journal in the same year. Then one clicks on the "Citation Report" button in WOS. This option sorts the papers by citation frequency and calculates the average citation rate. To calculate the PRI one needs the

citation rank of the paper and the number of papers in the year set of the journal. Both are provided by the "Citation Report" option.

$$PRI = (N - R + 1)/N*100,$$

where N is the number of papers in the year set of the journal, R is the descending citation rank of the paper (among the papers of the journal published in the year of the target paper). In case of ties (several papers having the same citation frequency), each of the tied values is assigned the average of the ranks for the tied set. Thus, if a target paper is the most cited paper in a journal in a year, its PRI = 100. Consider the paper in the first line of Table 4 (evolution, 1987). Its citation frequency (by January, 2009) is 313, which makes it the second most cited paper among 130 papers published in the journal Evolution in 1987. Thus, its PRI = (130-2+1)/130*100 = 99.2. We suggest that PRI values be rounded to whole numbers, and in this case, 99.

Table 5 gives data for the same author (as Table 4), but sorted by PRI. One can see that the actual number of cites differ dramatically for papers with the same PRI of 100: from 3 citations (fisheries, 2008) to 277 (j hered, 1998). Among the 29 papers ranked by PRI , there are 10 papers, which do not occur on the list of 29 most cited ones (contributing to the h-index), shown in Table 4. These ten papers are shown in Table 5 in bold face in the column labeled "Cites". We should stress, that very recent papers, not having yet accumulated many citations may attain a high PRI, like the paper in the journal "Fisheries", published in the last year (2008). For the majority of research fields, which are moderately to slow-moving ones, the citation number of 3 for a paper one year old is quite high as the average citation rate in January 2009 for an average paper of 2008 in this journal is only 0.10 (the journal being authoritative in the field of fisheries, though a narrowly specialized one).

Having characterized an author's individual papers we can now proceed to characterize the authors' overall citation status, which is the second step in our evaluatory procedure: obtaining the ASI. To illustrate this step we extracted citation data for three authors in population genetics. The comparative summary data are given in Table 6. Authors 1 and 2 are similar in citation performance. Author 3 is citation-wise less successful. One can see, that overall citation numbers and h-indexes do differentiate Author 3 from the other two: 872 vs. 3303 and 3007, and 19 vs. 31 and 29. However, citation data for recent papers, of 2004-2008 are not much different among these authors. The most informative characteristic distinguishing Author 3 from the other two seems to be the number of papers, for which the PRI equals or exceeds a certain threshold: 99, 95 or 75. These numbers are the ASIs. Authors 1 and 2 authored and co-authored 10 papers each, which have PRI $\geq$ 99, while Author 3 has published only 1 paper that ranks this high. Thus, ASI_{99} values for these authors are 10, 10 and 1. ASI_{95} and ASI_{75} for the three authors are 20, 21, 4 and 46, 36, 17, thus confirming lesser citation-wise success of the 3d author. For the period of 2004-2008 the same trend is seen: for the Authors 1 and 2 the ASI at all the 3 levels are higher than for the Author 3: ASI_{99} are 1, 3, and 0; ASI_{95} are 5, 8, and 1; and ASI_{75} are 9, 16, and 6.

Discussion

Radicchi et al. (2008) argue that the relative indicator $c_f = c/c_0$, where c is the number of citations articles get and c_0 is the average number of citations per article for the discipline (or for the year), is an unbiased indicator for citation performance across disciplines and years. Aside from a very complex problem of delineating a science field to obtain the normalizing value of c_0, the meaning of c_f seems insufficiently informative. In the last column of Table 4 and 5 we list c_f values for the papers of our target author. One can see that for all papers which ranked the top-most, correspondingly having PRI = 100, c_f values range from 6.8 to 33.3. Thus, the value of c_f does not characterize the citation status of the paper unequivocally, whereas the PRI does. Considerable variation of c_f values for papers of a specified PRI value is partly due to the

skewness of citation distribution. If the PRI would have been a normally distributed variable, the paper with $c_f = 1$ would have PRI = 50. Consider the paper (evolution, 1989; Table 4, line 14). Its c_f is only 0.6, but its PRI = 71. At the same time for the paper (mar biol, 1984; Table 4, line 24) c_f is higher, 1.0, but PRI (= 63) is less than for the "evolution" paper. This is because the top-most paper in the journal (Rice, 1989) was extremely highly cited for that journal (Evolution), more than 6,000 times, which led to smaller c_f because of inflated $c_0 = 112.8$ (compare with $c_0 = 56.2$ for the same journal, but in 1987). Coefficients of correlation between PRI and c_f values for our exemplary 3 authors are rather low and range 0.50 to 0.68 (see Table 6).

Another advantage of our PRI-metrics (compared to raw citation frequencies or citation ranks) is in its recognition of size differences of journals and the age of the publication. For instance, the paper (j hered, 1990; Table 4, 17th line) was cited 66 times, which ranks it as the 6th. The paper in the 11th line (aquaculture, 1992) was cited 79 times, which also ranks as the 6th. As there are 358 papers in Aquaculture in 1992 this 6th rank corresponds to PRI = 99, while the same rank of 6 for the paper (j hered, 1990) corresponds to PRI = 95 because there are only 111 papers in the Journal of Heredity in 1990. PRI also accounts for the age of the paper. For instance, consider two papers published in the journal "trends ecol evol" in 2005 and 2007 (see Table 5, lines 11 and 13). The paper, published in 2005 got 111 cites, while the 2007 paper was cited only 25 times. But this value translates into PRI = 98, equal to the paper of 2005, cited 111 times.

It is often considered controversial to use journal impact factors to characterize the status of a paper. For papers that are not yet published but already accepted for publication this seems a reasonable interim estimate since there is no citation data available yet. So the reputation of the journal seems to be the only evidence of the putative impact or quality of the paper. For most papers published very recently this IF-wise approach to evaluation may be warranted since it takes some time for most papers to accumulate citations. Otherwise, it is certainly unwise to consider journal impact factor as a surrogate for the actual citedness of a paper, as it has been shown by many others that papers published in the same year in the same journal may be cited quite differently. Even in journals with a very high impact factor there may be many uncited papers. The PRI is free from indirect assessment. A paper may be at the 95 to 100 percentile rank irrespective of the journal impact factor. This might be considered a drawback: a mediocre paper among poor ones published in an obscure journal obtains a high PRI. Though, on the contrary, publication of a paper in high impact journal may cause an inflation of citation through a sort of hitch-hiking effect: an inflated citation score of a paper may be due to visibility and availability of the journal in which it is published rather than by its own merits.

PRI might be especially useful in humanities and social sciences, where publications in books (proceedings volumes, topical collections of papers, etc.) rather than in journals are more common than in natural sciences. In some fields of the latter like zoology, botany, geography, or geology publications of papers in books (collected articles) are also quite common.

It might prove useful to apply some thresholds: for instance to ignore journals and books containing less than K papers, or to disregard those authors' papers that were cited less than N times. Concrete values of K and N may differ in various cases or circumstances. It seems useful to subtract self-citations from the citation numbers, as when they are small (in cases of recent papers, or in fields with a low average citation rate) self-citations may constitute a substantial share of citations. The importance of exclusion of self-citations in calculating h-index is considered by quite a few authors (Schreiber, 2007; Engqvist & Frommen, 2008; Zhivotovsky & Krutovsky, 2008).

As was said above the numbers of papers with PRI equal to or higher than a specified value (99 or 95) distinguish the 3 authors summarized in Table 6. We called this characteristic the Author

Superiority Index (ASI_{99} and ASI_{95}). Possibly, for younger scientists, who are not yet experienced enough in writing first rate papers, or in consideration of grant applications to some regional sponsoring foundations, where the competition is not too strong, the number of papers at 75^{th} percentile (that is ASI_{75}) would be more informative. On the contrary, in situations of strong competition among mature scientists a more appropriate indicator would be the number of papers with PRI $\geq$ 99 (ASI_{99}).

The most important advantage of the suggested assessment method seems its independence from the subject field: a target paper is ranked among the papers of the same journal. It is understood that the journal should be a specialized one dedicated to a well-defined science field. Only in this case PRI and ASI would be specialty independent. Then, there is a problem with multidisciplinary journals, among which there are some very important, high-impact journals, such as Science, Nature, PNAS, and some other journals. PRI values calculated using these journals will be dependent on the specialty of the target papers: papers belonging to fast moving fields would certainly be ranked higher than papers in slow-moving fields. Though one should have in mind that the proportion of papers published in these high-impact journals is very small compared to the number of papers published in specialty journals. Thus, the specialty bias introduced by using the multidisciplinary journals should be small. The majority of scientists who undergo evaluation procedures would most probably have no papers published in Nature or Science. If a target person would happen to author (or co-author) a paper in these journals, it could be specially noticed by the evaluatory committee.

In calculating the PRI (and then obtaining the ASI) we used Thomson's Web of Science (WOS). Any other database, which provides comparable citation data, could be used as well, for example Elsevier's SCOPUS (www.scopus.com). It is also possible to obtain the PRI for papers published in sources (journals or books), not covered by any database. In this case it would be necessary to obtain separately citation data for all the papers published in this source in the same year as the target paper by consulting an appropriate database or the web. One would then use an electronic spreadsheet to sort the obtained citation numbers in descending order to find the citation rank of the target paper.

It is of course more time consuming to calculate the PRI than the h-index. However, by using the "Citation Report" option in the WOS database it is quite manageable. Considering the potential impact of these citation based methods on individual careers any scientist or evaluator can afford the time to obtain a result which is more relevant than a quick and dirty approach to evaluation. Hopefully, Thomson Scientific (or some other database) could add calculation of the PRI and ASI to its "Citation Report" at some time in the future.

To summarize: the authors suggested two indexes, 1) Percentile Rank Index (PRI), which shows the citation status of a paper, and 2) Author Superiority Index (ASI), which is based on PRI values of all the papers of the author under evaluation and shows the number of papers published by the author, for which PRI values are equal to or exceed a specified percentile rank (99, or 95, or 75). ASI allows for the evaluation and comparison of authors across different specialties, across different time periods, taking into account quite recent publications. The PRI is a novel measure of the citation-wise success of individual papers. It immediately shows the impact status of an author among his/her peers (the authors of papers in the topical journal, where the target paper is published) and judged by the peers (the scientists who read the topical journal and cite the target paper). The ASI based on PRI values of all the papers of the target author is a summary of judgments of the colleagues on the importance of the author's contributions. As PRI values weakly correlate both with citation numbers for a paper, its "raw" citation rank and c_f values (see Table 6), the suggested indexes, PRI and ASI are aimed to complement rather than substitute h-index and other citation indexes.

Cited Literature:

Bornmann, L. & Daniel, H.D. (2009). The state of h index research. Is the h index the ideal way to measure research performance? EMBO Reports, 10, 2-6.

Engqvist, L. & Frommen, J.G. (2008). The h-index and self-citations. Trends in Ecology and Evolution, 23, 250-252.

Garfield, E. (2008). http://garfield.library.upenn.edu/classics.html .

Hirsch, J.E. (2005). An index to quantify an individual's scientific output. Proc Natl Acad Sci USA, 102, 16569-16572.

Hirsch, J.E. (2007). Does the h index have predictive power? Proc Natl Acad Sci USA, 104, 19193-19198.

Pudovkin, A.I. & Garfield, E. (2004). Rank-normalized impact factor: a way to compare journal performance across subject categories. Proceeding of the 67th ASIS&T Annual Meeting. Providence, RI, 41, 507-515.

Radicchi, F. & Fortunato, S., & Castelano, C. (2008). Universality of citation distributions: Toward an objective measure of scientific impact. Proc Natl Acad Sci USA, 105, 17268-17272.

Rice, W.R. (1989). Analyzing tables of statistical tests. Evolution, 43, 223-225.

Schreiber, M. (2007). Self-citation corrections for the Hirsch index. EPL, 78 (3), Article Number 30002.

Zhivotovsky, L.A. & Krutovsky, K.V. (2008). Self-citation can inflate h-index. Scientometrics, 77, 373-375.

Table 1.
Citation Numbers of Two Authors with the Same h-index but different overall citation numbers

Paper citation rank	Author A # of cites	Author B # of cites
1	21	195
2	11	57
3	10	45
4	10	30
5	10	25
6	9	18
7	8	12
8*	**8**	**9**
9	6	7
Total Cites	93	398

*8 = h-index.

Table 2.

Paper Citation Rates Depend on Paper Age, A and Science Field, B (ISI Essential Science Indicators, 1999-2008)

A			B	
Publication Year	**Av. Citation All Fields**		**Science Field**	**Av. Citation 1999-2008**
1999	17.69		Molecular Biology & Genetics	23.90
2004	10.63		Plant & Animal Science	6.83
2008	0.98		Mathematics	3.00

Table 3.
Medians and Quartiles of Journal Impact Factors in 3 Science Fields (ISI Journal Citation Report, 2008)

Science Field (JCR Category)	Number of Journals	Median IF	1st Quartile	3rd Quartile
Biochemistry & Molecular Biology	275	2.624	1.480	4.311
Zoology	125	1.072	0.648	1.612
Mathematics	215	0.562	0.421	0.826

Table 4.
Citation data and calculation of Percentile Rank Index. Shown are 29 papers, which are the h-core papers for this author. The papers are ranked by the numbers of cites.

c_0 is the average citation rate for the papers of the source (journal or book) by January 9, 2009.
$c_f = c/c_0$ (c, c_0 and c_f are the notations used in Radicchi, 2008)

Paper citation rank	Journal Title	Year	Cites received by the paper(c)	Papers in the journal (N)	Citation Rank within the source (R)	c_0	PRI	c_f
1	evolution	1987	**313**	130	2	56.2	99	5.6
2	genetics	1989	**287**	381	8	62.3	98	4.6
3	j hered	1998	**277**	105	1	22.8	100	12.2
4	can j fish aquat sci	1991	**169**	325	5	31.1	99	5.4
5	evolut aquat ecol	1995	**154**	38	1	22.8	100	6.8
6	pacific sci	1982	**128**	62	1	9.0	100	14.3
7	trends ecol evolut	2005	**111**	130	4	27.7	98	4.0
8	mol ecol	2006	**106**	353	1	10.4	100	10.2
9	fisheries	1999	**82**	155	3	5.4	99	15.1
10	cons biol	1990	**81**	66	10	44.2	86	1.8
11	aquaculture	1992	**79**	358	6	16.2	99	4.9
12	trends ecol evolut	2004	**74**	134	20	34.5	86	2.1
13	copeia	1986	**73**	149	7	17.5	96	4.2
14	evolution	1989	**72**	164	48	112.8	71	0.6
15	j hered	1990	**71**	111	5	16.4	96	4.3
16	can j fish aquat sci	1994	**70**	324	18	26.5	95	2.6
17	j hered	1990	**66**	111	6	6.4	95	4.0
18	trends ecol evolut	2003	**57**	137	39.5	46.4	72	1.2
19	can j fish aquat sci	1990	**56**	290	38	30.9	87	1.8
20	genetics	1988	**53**	284	86.5	56.1	70	0.9
21	can j fish aquat sci	1990	**46**	290	51	30.9	83	1.5
22	evolution	2004	**41**	270	15	18.8	95	2.2
23	fish bull	1987	**38**	77	14	21.3	83	1.8
24	mar biol	1984	**35**	240	90.5	35.6	63	1.0
25	cons biol	1998	**34**	192	69	38.0	65	0.9
26	t am fish	1996	**33**	107	13	17.9	89	1.8
27	genetics	2002	**33**	482	127	25.4	74	1.3
28	cons biol	2002	**32**	196	51.5	26.2	74	1.2
29 = h	**deep-sea res**	**1983**	**30**	**85**	**30.5**	**32.5**	**65**	**0.9**

Table 5.
Citation data and calculation of Percentile Rank Index. Shown are 29 papers, which are the h-core papers. The papers are ranked by the PRI.

For c, c_0, c_f see Table 2.

Paper rank by PRI	Journal Title	Year	Cites received by the paper (c)	Papers in the journal (N)	Citation Rank within the source (R)	c_0	PRI	c_f
3	fisheries	2008	**3**	87	1	0.1	**100**	33.3
3	mol ecol	2006	106	353	1	10.4	**100**	10.2
3	j hered	1998	277	105	1	22.8	**100**	12.2
3	ev aq eco syst	1995	154	38	1	22.8	**100**	6.8
3	pacific sci	1982	128	62	1	9.0	**100**	14.3
8	Evolution	1987	313	130	2	56.2	**99**	5.6
8	mol ecol res	2008	**3**	385	5	0.2	**99**	13.0
8	can j fish aq sci	1991	169	325	5	31.1	**99**	5.4
8	fisheries	1999	82	155	3	5.4	**99**	15.1
8	aquacult	1992	79	358	6	16.2	**99**	4.9
12	trends ecol evol	2007	**25**	115	3	7.3	**98**	3.4
12	genetics	1989	287	381	8	62.3	**98**	4.6
12	trends ecol evol	2005	111	130	4	27.7	**98**	4.0
14.5	cons genet	2006	**11**	96	4	3.6	**97**	3.0
14.5	fish fisheries	2008	**5**	29	2	1.2	**97**	4.3
17	j hered	1990	71	111	5	16.4	**96**	4.3
17	copeia	1986	73	149	7	17.5	**96**	4.2
17	fisheries	1990	**29**	48	3	5.1	**96**	5.7
20	j hered	1990	66	111	6	16.4	**95**	4.0
20	evolution	2004	41	270	15	18.8	**95**	2.2
20	can j fish aq sci	1994	70	324	18	26.5	**95**	2.6
22	j fish biol	2001	28	316	22.5	12.8	**93**	2.2
23.5	genetics	2007	**9**	615	56.5	3.8	**91**	2.4
23.5	mol ecol	2007	**9**	430	40	4.3	**91**	2.1
25.5	ecol gen impl	2007	**1**	27	4	0.3	**89**	3.8
25.5	t am fish	1996	33	107	13	17.9	**89**	1.8
27	pacific sci	1981	**17**	41	6	10.1	**88**	1.7
28	can j fish aq sci	1990	56	290	38	30.9	**87**	1.8
29	cons biol	1990	81	66	10	44.2	**86**	1.8

Table 6.
Comparative citation statistics for 3 authors in population/evolutionary genetics.

	Author 1	Author 2	Author 3
Data for all years			
Years	1973-2008	1981-2008	1976-2008
No of papers	92	67	66
Sum of cites	3303	3007	872
Av.cites/paper	35.90	44.88	13.21
h-index	31	29	19
No. of papers with			
PRI $\geq$ 99 (ASI_{99})	10	10	1
PRI $\geq$ 95 (ASI_{95})	20	21	4
PRI $\geq$ 75 (ASI_{75})	46	36	17
Correlation between			
Cites/PRI	0.53	0.59	0.55
Paper rank/PRI	- 0.48	- 0.38	- 0.47
PRI/av.cites	0.62	0.50	0.68
Data for 2004-2008			
No of papers	20	26	27
Sum of cites	107	445	96
Av.cites/paper	5.35	17.16	3.56
h-index	6	9	6
No. of papers with			
PRI $\geq$ 99 (ASI_{99})	1	3	0
PRI $\geq$ 95 (ASI_{95})	5	8	1
PRI $\geq$ 75 (ASI_{75})	9	16	6