The Adapted Pure *h*-Index

Jing- chun Chai[1] Ping-huan Hua[1] Ronald Rousseau[3,4] Jin-kun Wan[1,2]

04 June 2008

Abstract

The pure *h*-index, introduced by Wan, Hua and Rousseau (2007) takes the number of collaborators, possibly the rank in the byline and the actual number of citations into account. We propose a new method to calculate an alternative *h*-type index, referred to as the adapted pure *h*-index. This method leads to a new Hirsch core. We claim that our new approach is an improvement over the pure *h*-index as it is less biased with respect to authors with many multi-authored articles.

1 Introduction

Since the *h*-index was proposed by J. E. Hirsch in 2005 (Hirsch, 2005), this index of citation excellence has become a focal point in international scientometrics (Ball, 2005, 2007; Perkel, 2005). The *h*-index has been accepted as an index providing a good representation of a scientist's lifetime achievement (Bornmann & Daniel, 2007). It has, moreover, been shown to provide an acceptable (best among a series of other single-number indicators) prediction for future achievements (Hirsch, 2007). Although introduced in the context of publications and citations, it became soon clear (Braun *et al*., 2006; Egghe & Rousseau; 2006) that the basic structural framework for its calculation can be applied to many other source-item relationships. Following this idea *h*-indices have been calculated for journals (Braun *et al.*, 2006), topics (Banks, 2006; Rousseau, 2007; STIMULATE, 2007), library loans (Liu & Rousseau; 2007) and over other publication-citation periods than a scientist's total career. For a review on recent developments related to the *h*-index we refer the reader to (Rousseau, 2008).

2 Variations on the *h*-index theme

Although clear and simple, or maybe because it is clear and simple, several colleagues have tried to improve the *h*-index. Many of these proposals do not use citations in order to determine a cut-off line but take the actual number of citations into account. Among these we mention Egghe's *g*-index (Egghe, 2006), Jin's *A*-index (Jin, 2006) and the *R*-index proposed by Jin *et al*. (2007). Other proposals deal with the calculation process in order to determine the cut-off line. Examples of this type of variations on the *h*-index theme are: calculating a real-valued or a rational *h*-index (Rousseau, 2006; Ruane & Tol; 2008), taking self-citations into account or not (Schreiber, 2007), or considering different databases (Bar-Ilan, 2008; Sanderson, 2008). In

[1] Library of Tsinghua University, Beijing 100084, China; wanjk@lib.tsinghua.edu.cn; cjc3495@cnki.net
[2] China Scientometrics and Bibliometrics Research Center, P.O. Box 84-48, Tsinghua University, Beijing, 100084, China; gfhs@cnki.net
[3] KHBO (Association K.U.Leuven), Industrial Sciences and Technology, B-8400 Oostende, Belgium; ronald.rousseau@khbo.be
[4] K.U.Leuven, Steunpunt O&O Indicatoren, Dekenstraat 2, B-3000, Leuven, Belgium

H. Kretschmer & F. Havemann (Eds.): Proceedings of WIS 2008, Berlin
Fourth International Conference on Webometrics, Informetrics and Scientometrics & Ninth COLLNET Meeting
Humboldt-Universität zu Berlin, Institute for Library and Information Science (IBI)

this list of different calculation procedures we now focus on the idea of taking the number of co-authors into account.

3 Taking co-authorship into account

In his original article Hirsch (2005) observed that the *h*-index is a better single-number indicator than ‘total number of citations received’ because the total number of citations can be inflated by a small number of hits, maybe even co-authored with many others. Yet, he does not discuss the fact that also the *h*-index can be inflated if a scientist has written many co-authored articles. Probably Batista *et al.* (2006) were the first to consider the idea of taking collaboration into account. They simply divide *h* by the average number of researchers in the publications of the Hirsch core, leading to an indicator denoted h_I. As an alternative they propose dividing by the median number of researchers. This alternative eliminates problems in case a researcher has been part of a large team (as is sometimes the case in particle physics, epidemiology or demography). Burrell (2007) noted that hyper-authorship, quite common in many fields, may lead to an inflated *h*-index for individuals. Clearly, credit should be discounted, and hence also the contributors’ *h*-index.

Recently Wan *et al.* (2007) introduced another way of taking the number of co-authors and – possibly - the rank of a given author, into account. Their so-called pure *h*-index, denoted as h_p, is defined as follows.

First a normalized score is determined for each author in a publication. The term ‘normalized’ refers to the fact that the sum of all scores of one publication must be one. Hence an author of a single-authored paper always receives a score of one for his contribution. If *N* co-authors receive an equal score, then this score must be *1/N*. However, other scoring methods, taking (unequal) contributions into account are allowed. Next the equivalent number of co-authors of author A in document D, denoted by N_E (A,D), is defined as $\frac{1}{S(A_D)}$, where $S(A_D)$ denotes the normalized score of author A in document D. The pure *h*-index of author A, denoted by $h_p(A)$ is then defined as

$$h_p(A) = h\sqrt{\frac{h}{\sum_{D \in H(A)} N_E(A,D)}} \qquad (1)$$

where H(A) denotes the *h*-core of author A and the square root serves as a suitable sub-linear function applied to E(A) = $\frac{\sum_{D \in H(A)} N_E(A,D)}{h}$, the average equivalent number of authors of scientist A’s core articles. Clearly, when author A has written all his/her articles in the *h*-core as sole author, $h(A) = h_p(A)$. In all other cases $h_p(A) < h(A)$.

4 Taking co-authorship into account and changing the *h*-core

The previous approach never changes the original *h*-core. Yet, it might be argued that the approach outlined in (Wan *et al.*, 2007) is at times still biased in favour of authors with many collaborators. Consider for instance two authors each with an *h*-index equal to 10. The first one has an article that is cited 12 times, and this article is the result of collaboration among 10 authors, where author A is nor the leading nor the corresponding author. Author B has a single-authored article that is cited 9 times. Author A’s 10-person article contributes to his *h*-index and hence to his pure *h*-index, while author B’s single-authored article does not contribute to her *h*-index and hence also not to her pure *h*-index. Clearly, the pure *h*-index is still biased in favour of multi-authored papers, and hence, in favour of scientists that write many multi-authored papers. Yet, in order to correct for this bias one must be willing to change the original *h*-core. Indeed, Burrell (2007) already notes that if discounting is performed before the determination of the *h*-core this core itself can be altered.

Recently, Schreiber (2008) proposed a simple modification to the *h*-index calculation which takes co-authorship into account, and which changes the *h*-core. This approach has already been proposed by Egghe (2008). Indeed, Egghe (2008) presents a mathematical theory of the *h*-index (and also of the *g*-index) in case of fractional counting. In this theory Egghe considers fractional counting of citations as well as fractional counting of publications.

Fractional counting of publications is the proposal published by Schreiber. Articles are ranked according to the number of citations, but the rank itself grows according to the number of co-authors. This leads to an *h*-index, denoted as h_m, (and denoted h_F by Egghe) for which the h_m-core (the h_m-defining set of articles) contains more articles than the *h*-core. Yet, all articles belonging to the *h*-core also belong to the h_m-core.

5 An adapted pure *h*-index

In this article we propose the following method to calculate an adapted pure *h*-index. This new *h*-index will be denoted as h_{ap}. In order to determine this index for author A the following steps must be taken.

Step 1. Determine for each article co-authored by A the equivalent number of co-authors in the same way as for the pure *h*-index;
Step 2. Determine for each article the equivalent number of citations, defined as the actual number of citations obtained divided by the square root of the equivalent number of authors (of A);
Step 3. Rank A's articles according to the equivalent number of citations.
Step 4. As the equivalent number of citations is a real number, also h_{ap} is defined as a real number. This is done applying the method already proposed by Rousseau (2006) for the (real-valued) *h*-index. Let us denote the equivalent number of citations received by the article ranked *r* as $C_E(r)$, and its piecewise linear interpolation as $C_E(x)$, this is: the function connecting the points $(r, C_E(r))$, where *r* denotes the rank (*r* = 1, 2, ...), then h_{ap} is defined as the abscissa of the intersection of the lines *y* = *x* and the observed function $C_E(x)$. If the *h*-index of the equivalent number of citations is h_e then h_{ap} lies between h_e (inclusive) and h_e+1. It is easy to check that h_{ap} is obtained as:

$$h_{ap} = \frac{(h_e+1)*C_E(h_e) - h_e*C_E(h_e+1)}{C_E(h_e) - C_E(h_e+1)+1} \quad (2)$$

In particular, if $C_E(h_e) = C_E(h_e+1)$ then $h_{ap} = C_E(h_e)$. If the last rank is $r = h_e$ then the previous formula (2) can not be used. There are two options in this case: one either takes $h_{ap} = r$ or one adds a fictitious article with zero citations and one calculated h_{ap} using formula (2). If an author's articles are all single-authored, then her real-valued *h*-index is always equal to her $h_p = h_{ap}$. In general the h_{ap} is always smaller than or equal to the real-valued *h*-index. All articles having an equivalent number of citations larger than or equal to h_{ap} form the h_{ap}– core. As will be shown in the next (fictitious) examples, this h_{ap} – core can be totally unrelated to the *h*-core.

Our proposal does not coincides with Egghe's (2008), even if author's have an equal contribution. The difference is that Egghe divides citations by the number of authors, while we divide by the square root of the (equivalent) number of authors. Taking a square root reduces the influence of mega-authored articles on an author's h_{ap} -index.

6 Two fictitious examples

We present an extreme, fictitious, example where none of the articles in the new core belongs to the standard *h*-core (Table 1).

Table 1: Author A's most-cited articles (fictitious example)

A	B	C	D	E
Article 1	10	7	3.78	4
Article 2	10	7	3.78	5
Article 3	8	5	3.58	6
Article 4	8	5	3.58	7
Article 5	6	4	3	8
Article 6	6	4	3	9
Article 7	5	1	5	1
Article 8	5	1	5	2
Article 9	4	1	4	3
Article 10	4	2	2.83	10
Article 11	4	2	2.83	11

A: articles; B: number of citations; C: number of authors = equivalent number of authors; D:

equivalent number of citations; E: new rank (step 3)

Author A's *h*-index is 6. Assuming that the equivalent number of authors is always equal to the actual number of authors, his average equivalent number of authors is equal to 32/6, and hence his pure *h*-index is equal to 2.60. The new, adapted pure *h*-index is 3.82. The h_{ap} – core consists of the original articles 7, 8 and 9, none of which belonged to the original *h*-core.

We know consider another example where we take the rank of author B into account by using arithmetic counting (Table 2). Recall that in this counting procedure the author who is ranked R among N co-authors receives a normalized score of $\frac{2}{N}\left(1-\frac{R}{N+1}\right)$ and the corresponding equivalent number of authors of this article is $\frac{N(N+1)}{2(N+1-R)}$.

Table 2: Author B's most-cited articles (fictitious example)

A	B	C	D	E	F
Article 1	10	4th among 5	7.5	3.65	4
Article 2	10	4th among 4	10	3.16	6
Article 3	8	3rd among 3	6	3.27	5
Article 4	8	4th among 4	10	2.53	7
Article 5	6	3rd among 3	6	2.45	8
Article 6	6	4th among 4	10	1.90	10
Article 7	5	Sole author	1	5	1
Article 8	5	Sole author	1	5	2
Article 9	4	Sole author	1	4	3
Article 10	4	2nd among 3	3	2.31	9
Article 11	4	3rd among 3	6	1.63	11

A: articles; B: number of citations; C: author rank in the byline and number of authors; D: equivalent number of authors; E: equivalent number of citations; F: new rank (step 3)

Author B's *h*-index is also 6. His average equivalent number of authors is equal to 49.5/6, and his pure *h*-index is equal to 2.09. The new, adapted pure *h*-index is 3.74. The h_{ap}– core consists of the original articles 7, 8 and 9, none of which belonged to the original *h*-core.

7 Practical considerations

We admit that there is a precision problem and that the practical calculation of the h_{ap}-index is more difficult than the calculation of the *h*-index or the h_p-index. Yet, with a suitable software program this does not need to be a real problem. Indeed, we wrote a program suitable for use in the CNKI (China National Knowledge Infrastructure) database. This program collects and calculates the information shown in Table 3.

Table 3: Information obtained from our CNKI program

A	B	C	D	E	F	G
Author I	6	11	2.09	34.90	14.00	3.74
Author II	5	7	2.74	29.63	21.29	4.15
Author III	5	7	2.67	25.75	15.06	3.52
Author IV	5	5	2.11	20.52	14.15	3.88
Author V	2	2	1.07	8.71	8.71	2 (or 2.43)

A: Authors; B: *h*-index; C: number of articles; D: h_p; E: equivalent number of citations; F: equivalent number of citations of articles in the core; G: h_{ap}

Author V is a special case where $r = 2 = h_e$. Adding a fictitious article with zero citations leads to a h_{ap} value equal to 2.43.

8 A real-world example

As an example we determined Ronald Rousseau's h, h_p and h_{ap}-index based on Web of Science data (May 31, 2008). Rousseau's h-index is 16, his pure h-index is $h/\sqrt{2.4375} \approx 10.25$, and his adapted pure h-index is 12.43. His h_{ap}-core consists of 12 articles, three of which (single-authored) do not belong to his h-core.

9 Note

In practical situations the h-index is usually larger than or equal to h_{ap}. Yet, it is possible that $h < h_{ap}$. Consider, for instance, Table 4.

Table 4: Author C's most-cited articles (fictitious example)

Rank	Citations	Authors	Equiv. citations
1	5	4	2.50
2	3	1	3
3	2	1	2

Author C's h-index is 2. His h_e-value is also 2. His h_{ap}-value is: 2.33. Note that the original real-valued h is different from h_{ap} and equal to 2.5. Also when a fictitious article when zero citations is added it may happen that $h < h_{ap}$ (see Table 3).

10 Discussion

Many known h-type indices such as the h-index itself, the R-index (Jin *et al.*, 2007) and the pure h-index make use of the same core as the original h-index. The proposal presented in this contribution changes the core itself, adapting it to observed citation data. We think that our approach results in a less biased, hence more logical result.

It was already observed by (Wan. *et al.*, 2007) that the pure h-index could be increased by using another core. Yet, we considered this approach too time-consuming and dismissed it at that time. In the terminology of (Liu & Rousseau, 2007) the precision problem increased. In retrospect we think that fairness should prevail, and hence, although more tedious to determine, h_{ap} is to be preferred above h_p. As we consider the possibility that contributions are weighted, the h_{ap}-index is slightly more general than Egghe's fractional h-index, h_F. In general the precision problem can largely be ignored by using a dedicated software program similar to the one mentioned in this article for the CNKI database.

References

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

Ball, P. (2007). Achievement index climbs the ranks. *Nature*, 448, 737.

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

Bar-Ilan, J. (2008). Which h-index? A comparison of WoS, Scopus and Google Scholar. *Scientometrics,* 74, 257-271.

Batista, P. D., M. G. Campiteli, O. Kinouchi, and A. S. Martinez (2006). Is it possible to compare researchers with different scientific interests? *Scientometrics*, 68(1), 179-189.

Bornmann, L. and H.-D. Daniel (2007). What do we know about the h-index? *Journal of the American Society for Information Science and Technology*, 58(9), 1381-1385.

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

Burrell, Q.L. (2007). Should the h-index be discounted? In: W. Glänzel, A. Schubert and B. Schlemmer (eds.) *The multidimensional world of Tibor Braun*. Leuven: ISSI, pp. 65-68.

Egghe, L. (2006). An improvement of the h-index; the g-index. *ISSI Newsletter*, 2(1), 8-9.

Egghe, L. (2008). Mathematical theory of the h-index and g-index in case of fractional counting of authorship. *Journal of the American Society for Information Science* *and Technology*. To appear.

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

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.

Hirsch, J.E. (2007). Does the *h*-index have predictive power? *Proceedings of the National Academy of Sciences of the United States of America,* 104(49), 19193-19198.

Jin, BH. (2006). H-index: an evaluation indicator proposed by scientist. *Science Focus*, 1(1), 8-9.

Jin, BH., L. Liang, R. Rousseau and L. Egghe (2007). The *R*- and *AR*-indices: complementing the *h*-index. *Chinese Science Bulletin*, 52: 855-863.

Liu, YX and R. Rousseau (2007). Hirsch-type indices and library management: the case of Tongji University Library. In: *Proceedings of ISSI 2007* (Daniel Torres-Salinas & Henk F. Moed, eds.). Madrid: CINDOC-CSIC, pp. 514-522.

Perkel, J.M. (2005). The future of citation analysis. *The Scientist*, 19(20), 24.

Rousseau, R. (2006). Simple models and the corresponding *h*- and *g*-index. E-LIS: ID 6153.

Rousseau, R. (2007). Hungary – and Tibor Braun – on top! Dedicated to Tibor Braun on the occasion of his 75th birthday. In: W. Glänzel, A. Schubert and B. Schlemmer (eds.) *The multidimensional world of Tibor Braun*. Leuven: ISSI, pp. 23-26.

Rousseau, R. (2008). Reflections on recent developments of the *h*-index and *h*-type indices. *These Proceedings.*

Ruane, F.P and R.S.J. Tol (2008). Rational (successive) *h*-indices: an application to economics in the Republic of Ireland. *Scientometrics*, 75, 395-405.

Sanderson, M. (2008). Revisting h measured on UK LIS and IR academics. *Journal of the American Society for Information Science and Technolog,* 59(7),1184-1190.

Schreiber, M. (2007). Self-citation corrections for the Hirsch index. *Europhysics Letters*, 78, 30002.

Schreiber, M. (2008). To share the fame in a fair way, h_m modifies *h* for multi-authored manuscripts. *New Journal of Physics*, 10, 040201.

The STIMULATE-6 Group (2007). The Hirsch index applied to topics of interest to developing countries. *First Monday*, 12(2). http://www.firstmonday.org/issues/issue12_2/stimulate/

Wan, JK. PH. Hua and R. Rousseau (2007). The pure *h*-index: calculating an author's *h*-index by taking co-authors into account. *COLLNET Journal of Scientometrics and Information Management*, 1(2), 1- 5.