Qualitative and Quantitative Methods in Libraries (QQML) 5: 845-853, 2016

Towards designing the Indicator for Scholarly Academic Research Impact based on h-index

Sibsankar Jana

Asst. Professor, DLIS, University of Kalyani, E-mail: sibs_jana@yahoo.com

Abstract: Present paper discusses Scholarly Academic Research Impact (SARI). According to J. E. Hirsch, the main criteria for devising an indicator are number of citations, scientific age of author, subject of the papers, self-citations, simplicity, universal applicability, least sensitive to small change of bibliographic record etc. The h-index has included some of these criteria. However some criteria still need to be incorporated during calculating h-index. Here I have tried to consider some relevant components to h-index for better reflection of SARI. I think h-index will be in core but some other measurements should be incorporated for better reflection of the impact. The total number of citations or average number of citations per article should be included to for calculating the SARI. G-index includes some articles having very high citations but does not include citations beyond hirsch's core. But hT-index has the provision to give weightages of all the citations of all the articles even beyond Hirsch's core but it is very complicated to understand and to calculate. The scientist having higher scientific age naturally will get more citations than the younger, so we need to incorporate this criterion when calculating SARI. Therefore technique should be developed to conform the relation "SARI will be inversely proportional to Scientific Age of the researcher or scientist". The SARI should have universal applicability (for author, journal, publisher, country etc.). Again variations among the disciplines and subjects need to be kept in mind when designing SARI, so that the issue should be normalized. The language biasness is very hard to avoid during designing SARI. There are so many indicators like JIF, h-index and its variants, altmetrics etc, but my proposal is to develop a single indicator which should include all the aspects of all the indicators as far as possible. In the present context of electronic web environment we have to measure societal impact along with the scholarly academic impact. In this regard 'altmetrics' is basically the tool for weighing societal impact of the scholarly articles by counting number of downloaded, viewed, re-used, shared, liked etc. We cannot ignore these criteria in the social networking environment. But altmetrics is not so crystallized as it depends on web visibility and validation. Therefore present study intentionally has excluded altmetrics aspect of SARI. I have not given any concrete formula rather just have wanted to draw a theory for designing a SARI indicator based on h-index.

Received: 31.3.2016 Accepted: 21.6.2016 ISSN 2241-1925

Keywords: h-index; Scholarly Academic Research Impact (SARI); Tappered h-index; g-index; self-citation

1. Introduction

Any research has two dimensions of impact namely Academic Research Impact (ARI) and Social and Economic Research Impact (SERI).

A. Academic Research Impact (ARI): ARI is the impact of the research output on the academic domain. It has again two dimensions.

    1) Scholarly Academic Research Impact (SARI): Based on Citations
    2) Social Networking Academic Research Impact (SNARI): Based on Download, view, like, share etc.

B. Social and Economic Research Impact (SERI): SERI is the impact of research output on the society. It is very hard to measure.

The Social and Economic Research Impact is very hard to measure. The present study is mainly concerned with the SARI (Shaping Society). The impact of any research need to measure for the following reasons:

a) It enhance the self satisfaction of the researchers
b) It is used to assess for providing research grant to individual and organization
c) It helps to assess the grade of the institution / organization (e.g. NAAC grade)
d) It creates healthy competitive environment among authors and organizations
e) It is very much helpful to new researchers and students for identifying high impact research
f) It is used to prepare list for confirmation/promotion/yearly performance appraisal e.g. API

It is believed that number of citations got by a research publication is directly proportional to the academic research impact of that particular publication. Broadly, a citation is a reference to a published or unpublished source (not always the original source). More precisely, a citation is an abbreviated alphanumeric expression embedded in the body of an intellectual work. Generally, the combination of both the in-text citation and the bibliographic entry constitutes what is commonly thought of as a citation. The first instrument to measure the SARI is Journal Impact Factor (JIF).

The 2009 impact factor of a journal would be calculated as follows:

2009 impact factor = A/B

***A** = the number of times articles published in 2007 and 2008 were cited by journals, books, patent document, thesis, project reports, news papers, conference/ seminar proceedings, documents published in internet, notes and any other approved documents during 2009*

***B** = the total number of "citable items" published by that journal in 2007 and 2008. ("Citable items" are usually articles, reviews, proceedings, notes or any other documents pre-reviewed before publishing it)*

It is basically an indirect measure of a research publication. Still now, though the JIF is the popular way of weighing the research work or output, it has some questionable aspects. These are:

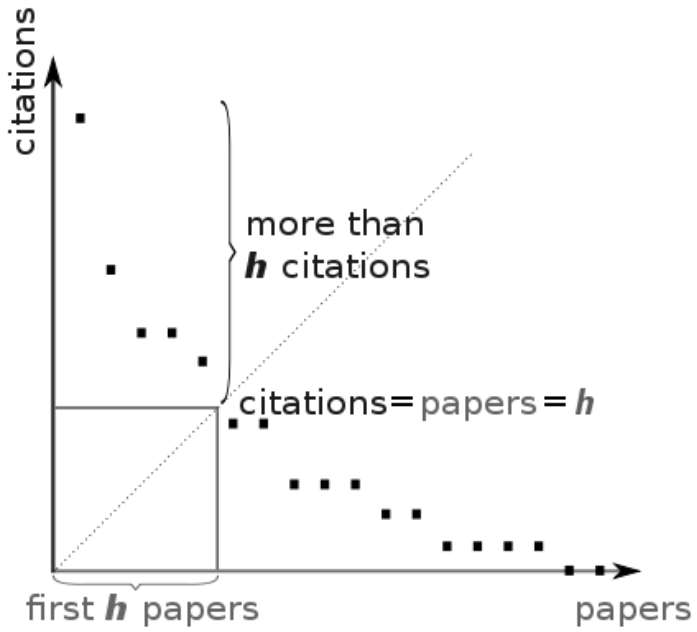
- **Discipline-wise variation:** The author cites recent articles in case of science and technology than social science and humanities. Therefore journal on science and technology automatically receive high JIF than others.
- **Citation bias may exist**: for example, English language resources may be favored.
- **Self-citation:** Authors may cite their own work.
- **Limited number of articles**: Not all research output are published and cited in the citation indexes, so uncounted during calculation of JIF.
- **Publication exclusion**: only research articles, technical notes and reviews are citable items. Editorials, letters, news items and meeting abstracts are non-citable items.
- **Homograph problem:** It may be difficult to separate citations to two unrelated scientists who happen to share the same last name and first initial.
- **Citation score influences:** Too many un-cited or low-cited articles may enjoy high impact value due to a few highly cited articles.
- **Review articles**: Authors and journals that frequently publish review articles tend to have their citation counts more as these are usually highly cited items.
- **Cronyism:** friends or colleagues may reciprocally cite articles of each other to mutually build their citation counts.
- **Coercive citation**: The citations included by the author on request of an editor of a scientific or academic journal. Sometimes they force an author to add spurious citations to an article before the journal will agree to publish it.

2. Factors to be Considered for Designing SARI

The h-index was proposed by J.E. Hirsch (a physicist at the University of California) in his paper "An index to quantify an individual's scientific research output". A scientist has index h if h of his/her Np papers have at least h citations each, and the other (Np-h) papers have no more than h citations each (Schubert & Glänzel, 2007). It is a simple metric which quantify the output of an

individual researcher. For example if one has published 10 papers that have received at least 10 citations each then h-index is 10.

Figure-1: Graph representing h-index



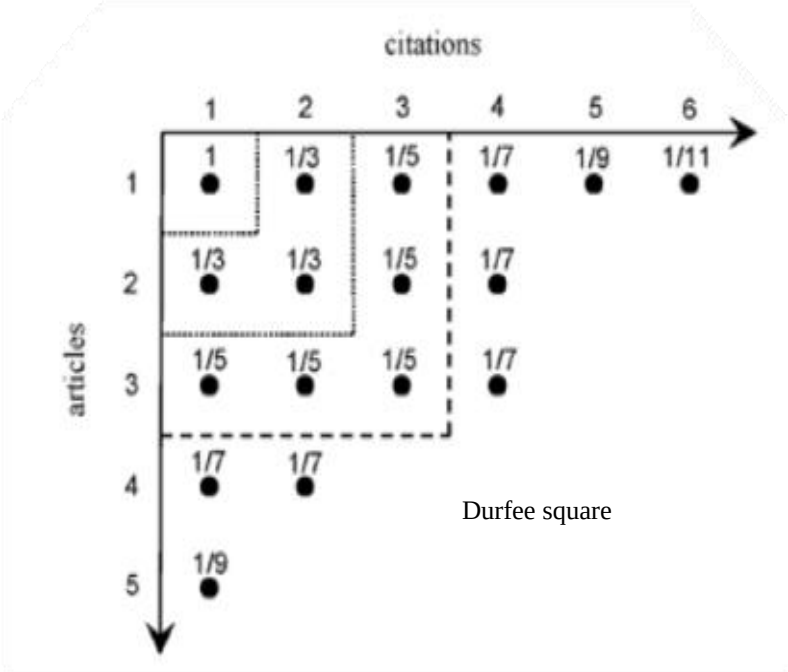
The main idea behind the emergence of h-index according to J.E.Hirsch is "the number of citations received by a scientist's publications is a better indicator of the quality of the scientist than the number of papers published or the journals where they were published"(Hiresch & Buela-Casal, 2014).
The points to be considered for designing impact indicator are (Hirsch, 2007):

- Number of citations received by the articles of an author
- Self-citation
- Easy to determine i.e Simplicity
- Universal applicability
- Scientific age of the researcher
- Least sensitive to small change of bibliographic record

2.1 Number of citations received by the article(s) published by an author
In h-index, citations are counted only for the articles, which are in the Hirsch's core. The most productive articles represent the h-index of the author. But in case of some articles which got exceptionally high citations are not reflected by

h-index. There should be some provisions to take into account the total number of citations received by the articles published by an author in addition to the h number of articles having h number of citation each. The g-index was proposed by Leo Egghe to overcome this problem. It aims to improve on the h-index by giving more weight to highly-cited articles. "Given a set of articles ranked in decreasing order of the number of citations that they received, the g-index is the (unique) largest number such that the top g articles received (together) at least g^2 citations." (Egghe, 2006) As for example: an academic has a g-index of 30 if the top 30 most cited of his/her papers combined have at least 30^2 i.e. 900 citations. The e-index "Defined as the square root of the surplus of citations in the h-set beyond h^2, that is, beyond the theoretical minimum required to obtain a h-index of h". If an academic has an h-index of 10, but has a total of 1000 citations to the first 10 published articles, his/her e-index would be 30 (the square root of 1000 minus the theoretical minimum required to obtain a h-index of 10, i.e. 100). Therefore e-index will be √(1000-100) equals to 30. The aim is to differentiate between scientists with similar h-indices but different citation patterns. It gives more attention to highly-cited articles. But it does not take into account the citations beyond Hirsch's core.

Figure-2: Ferrers graph



In Tappered h-index (hT-index) all the citations have the power to influence the impact. If we take an example it will be very much cleared. A scientist who has 5 publications which, when ranked, have 6,4,4,2,1 citations. This publication output can be represented by a Ferrers graph (Figure-2). The largest completed

(filled in) square of points in the upper left hand corner of a Ferrers graph is called the Durfee square. The h-index is equal to the length of the side of the Durfee square (in this case h = 3) (Anderson, Hankin, & Killworth, 2008).

h(1) = 1 citation
h(2) = 4 = h(1) + 3 additional citations
h(3) = 9 = h(2) + 5 additional citations
h(4) = 16 = h(3) + 7 additional citations
And so on…
h-index = Scores in Durfee square = (1 + 1/3 + 1/3 + 1/3 + 1/5 + 1/5 + 1/5 + 1/5 + 1/5) = 3
hT -index = h-index + other scores outside Durfee square = (3 + 1/7 + 1/7 + 1/7 + 1/7 + 1/7 + 1/79+ 1/9 + 1/11) = 4.03

hT-index scientifically considers all the citations (within and outside Hirsch's core) received by all the articles of an author. But it is too complicated to understand and to measure the impact.

My suggestion:

a) In case of shared responsibility, the citations should be divided among authors before counting the number of citations.
b) The citations beyond Hirsch's Core should be taken into account, such that SARI= √(h x Average citations per article). Example
    - ❖ In ideal situation i.e. if an author got 10 citation each for 10 publications, get SARI= √(h x 10)= √(10x10)=10=h-index.
    - ❖ But if he get average citations less than h then SARI <h-index
    - ❖ But if he get average citations greater than h then SARI >h-index

2.2 Self-Citation

A self citation $C^{S}_{P\rightarrow Q}$ is any citation appearing in a paper P pointing to paper Q, whose set of authors are respectively A[P] and A[Q], for which it holds true: $A[P] \cap A[Q] \neq \emptyset$, i.e., the intersection of the sets of authors is not empty. Whenever there is talk about citation analysis, the first concern always seems to be to exclude self-citations. It is generally believed that authors can inflate their citations by excessive and intentional self-citations. Though, there are so many opinions regarding self-citations.

My Suggestion:
In the calculation of h-index, self-citations can artificially increase the value of h-index of an author**.** I think self-citations should be avoided during calculation of ARI.

2.3 Easy to determine i.e Simplicity

The indicator must have the quality of simplicity in regards to the calculation and understanding. The h-index, in this respect is very much simple to

understand and easy to calculate. Many software or online services like Bibexcel, Publish or Perish, Google Scholar are used to measure h-index. But over simplicity sometimes may not represent all the criteria exhaustively. In case of h-index, the scientific age of the author and citations beyond Hirsch's core are not considered

2.4 Universal Applicability

The indicator must have the quality of universal applicability.

a) It should be used for author, journal, publisher, country irrespective of language, subject, time etc.
b) JIF is only applicable for journal. We can measure h-index of an author, a journal, a publisher, a country etc.
c) But the visibility of citing documents can play important role during h-index calculation. Suppose document A (cited document) is cited by document B (citing document), but B is not visible and therefore not consider for h-index calculation.
d) Again, in general, h-index significantly varies from discipline to discipline and subject to subject. It is higher in science than humanities and social science with respect to time. Therefore, only authors of same subject/discipline are to be compared.
e) Languages of the publications also have great impact on citations. Articles published in English language get more citations than others.

My Suggestion:

a) Average time (years) taken for 1st citation of articles (AvgC1) of different discipline have to be taken into account. E.g. If AvgC1 for Humanities is 4 years and if an article get citation within 1 year of publication then citation count will be 4. For citation on 4th year will get citation count 1. For citation on 6th year will get citation count 4/6. In case of Science, obviously the AvgC1 will be less than Humanities. In this way subject biasness should be normalized. For this an exhaustive study should be required for calculating AvgC1 of each discipline/subject. Then we can normalize the citation counts of an article from discipline-wise biasness and publication-wise biasness.
b) Device should be developed to minimize the language biasness. The citation weightage should be inversely proportional to the Scientific population of the particular language. In case of English, the Scientific population is higher so citation weightage will be less. It is believed that articles have the possibility of getting more citations. 2 citations of an article of Bengali language may be equivalent to the 6 citations of English.

2.5 Scientific age of the researcher

The h-index of an author increases with "scientific age" (i.e., the time elapsed since their first publication). The younger authors obviously have comparatively low h-index than aged authors. In this context, we may use the concept of "m

index" (the quotient of the h-index divided by the number of years elapsed since the scientist's first publication) as a "timeless" index to make comparisons between scientists who are at different stages of their career.

My Suggestion:
To normalize the SARI the Scientific age of the researcher need to be taken into account. It will indicate the SARI curve through out the career of a researcher.

2.6 Least sensitive to small change of bibliographic record

Indicator should be like that any little bit change in bibliographic record (number of citations, scientific age of the author etc) does not responsible to great change on h-index. The h-index is very least sensitive to any change in bibliographic record. Mare increase of few citations does not affect the overall h-index. It is the beauty of h-index.

3 Basic framework for designing a new indicator

In my opinion, the new indicator is required to devise that should includes the following aspects one by one:

a) Self-citation if any should be avoided first.
b) In case of shared responsibility, total citation got by an article will be divided by number of authors and the resulted citation count equal to or greater than 0.5 will be treated as 1.
c) Citation counts should be normalized by avoiding language and subject biasness.
d) Now we will get standard revised citation counts for each article of an author.
e) Now we calculate h-index on the basis of revised citation count.
f) Then this h-index is multiplied by average citation count of the article. Then we find square root of this result. It gives new h-index.
g) At last this new h-index is divided by the Scientific age of the author and we will ultimately get the revised h-index. We may call it Revised h-index or **h_r-index.**

$$hr - index = \frac{\sqrt{(h \times C)/N}}{T}$$

Where,

h = h-index on the basis of revised citation
C = Total number of citation of all articles of an author
N = Total number of articles of an author
T = Scientific age of the author (in years)

4 Further Research

In the present context of electronic information environment, altmetrics i.e. article-level metrics are really very useful for calculating the impact of any

publication. Altmetrics cover not just citation count, but also other aspects of the impact of work, such as how many times data and knowledge bases, article ,websites, software, blog, slides, figure, video, audio has been viewed, shared, downloaded, reused, liked, cited etc. In growing numbers, scholars are moving their everyday work to the web. This matters because expressions of scholarship are becoming more diverse. Articles are increasingly joined by:

a) The sharing of "raw science" like datasets, code, and experimental designs
b) Semantic publishing or "nanopublication," where the citable unit is an argument or passage rather than entire article.
c) Widespread self-publishing via blogging, microblogging, and comments or annotations on existing work.

Still no single indicator like h-index, has not yet been crystallized till date in altmetrics. Only we can measure number of liked, shared, downloaded, viewed etc. Again two important issues arise in altmetrics are visibility and validity. All the articles of an author are not in digital form so not visible to others and hence have no chance of download, like, etc. On the other hand, who will validate the number of altmetrics measures? Anyone can easily inflate the numbers by self-downloading, self-sharing, self-liking etc.

Again, we can devise a new indicator having two components viz. citation and download/view/like/share etc. Again further study may be conducted to find out any correlation between numbers of citations to number of download/view/share/like etc. Then we may say that:

A no. of Citation = B no. of Download = C no. of Share = D no. of Abstract view = E no. of Like and so on. Then, we can assign different points for citation, download, view, like etc. By adding these points, we can determine the total points of an article. At last, we can apply the analogy of h_r-index to measure the h-point.

References

Anderson, T., Hankin, K., and Killworth, P., (2008). Beyond the Durfee square: Enhancing the h-index to score total publication output. *Scientometrics* , Vol.76, No.3, 577-588.

Egghe, L., (2006). Theory and practice of the g-index. *Scientometrics* , Vol.69, No.1, 131-152.

Global Institute for Scientific Information., (n.d.). *Journal Impact Factor (JIF)*. Retrieved March 22, 2015, from http://www.jifactor.com/about.asp.

Hiresch, J., and Buela-Casal, G., (2014). The meaning of the h-index. *International Journal of Clinical and Health Psychology*, 14, 161-164.

Hirsch, J., (2007). Does the h-index have predictive power? *PNAS* , 104, 19193-19198.

Schubert, A., and Glänzel, W., (2007). A systematic analysis of Hirsch-type indices for journals . *Journal of Informetrics*, Vol.1, No.3, 179-184.

Shaping Society. (n.d.). Retrieved April 11, 2015, from http://www.esrc.ac.uk/funding-and-guidance/impact-toolkit/what-how-and-why/what-is-research-impact.aspx