SHORT COMMUNICATION

On the relation between Schubert's h-index of a single paper and its total number of received citations

L. Egghe

Universiteit Hasselt (Uhasselt), Campus Diepenbeek, Agoralaan, B-3590 Diepenbeek, Belgium

leo.egghe@uhasselt.be

ABSTRACT

A relation, established by András Schubert [Using the h-index for assessing single publications. Scientometrics 78(3),559-565, 2009] on the relation between a paper's h-index and its total number of received citations, is explained. The relation is a concavely increasing power law and is explained based on the Lotkaian model for the h-index, proved by Egghe and Rousseau.

I. Introduction

A remarkable new application of the h-index has been proposed in Schubert (2009) for assessing single publications: the h-index of a single article. It is defined as the h-index of the set of papers that cite this single article. In other words, fixing a single article, one looks at all articles that cite this article. In this set of citing articles, for each citing article, one looks at the number of citations to this citing article. To avoid confusion, we will call the single article as being on level 1. The citing articles to this single article are then on level 2. The articles that cite these articles on level 2 are considered to be on level 3.

If we rank the articles on level 2 in decreasing order of received citations (by articles on level 3), we can apply the definition of the h-index of this ranked situation: it is the highest rank r = h such that the articles on level 2 at ranks r = 1,...,h receive h or more citations (of articles at level 3). This is the classical definition of Hirsch of his h-index (Hirsch (2005)) (here applied to our case) and is applicable to any source-item situation as re-described below (see also Egghe (2005, 2009)): in any system where we have "sources" having (or producing) "items" (called an information production process (IPP)), we can rank the sources in decreasing order of the number of items they have. The h-index of this IPP is then the highest rank r = h such that the sources on ranks r = 1,...,h have h or more items.

The original definition of Hirsch was applied to sources = articles (of an author) receiving citations = items. Replacing articles of an author by articles in a journal we have the journal's h-index, introduced by Braun, Glänzel and Schubert (2006). The h-index of a topic was introduced by Banks (2006) (papers on a topic as sources and citations to these papers as items). We refer to Egghe (2009) for a comprehensive review (up to, and including 2008) of applications to other fields and further studies of the h-index and other h-type indices.

The application given in Schubert (2009) is special and unique since here, we think for the first time, citing articles (level 2, citing a single article on level 1) are considered as sources and the articles on level 3 (citing these articles on level 2) are considered as items.

In Schubert (2009) on finds, experimentally, for the articles (of level 1) in a journal (Schubert considers two journals, Cell and JACS), a relation between the articles' h-index (as described above) and their total number of received citations. It is explicitely stated in Schubert (2009) (last line of page 561) that "Semilogarithmic plot is used for more compact presentation, and a logarithmic fit is given to guide the eye, but no theoretical model suggesting such a functional relation is advised. (A power function also gives a tolerable fit)".

In the next short section, we will present a rationale for the latter regularity (a power law), based on our interpretation of the source-item situation applied here and based on the mathematical model for the h-index in a Lotkaian framework as presented in Egghe and Rousseau (2006). This will explain the regularity found in Schubert (2009).

II. Explanation of the functional relation between a single paper's h-index and its total number of received citations

In general IPPs, where we have sources having (or producing) items, we can define the h-index as indicated in the Introduction. Suppose now that sources have items according to the law of Lotka (Lotka (1926), Egghe (2005)). That means that the number of sources with n items is given by

$$f(n) = \frac{C}{n^{\alpha}} \tag{1}$$

where $n = 1, 2, ...$ and C > 0 and $\alpha > 1$ are parameters. For calculatory reasons one always uses (1) with continuous variable $n^{3}\ 1$ (Egghe (2005)). Under these conditions, it was proved in Egghe and Rousseau (2006) that the h-index of such a system is given by

$$h = T^{\frac{1}{\alpha}} \tag{2}$$

where T denotes the total number of sources. Note the crucial role of Lotka's exponent α here.

Now let us go back to the framework discussed in this paper: the h-index of a single paper (level 1) as introduced in Schubert (2009). Since this h-index is the h-index of the set of citing papers on level 2, calculated on the ranked list (in decreasing order) of these papers and on their received citations (from papers on level 3), we have here that T, the total number of sources, is the total number of citations (to the single paper on level 1) given by the citing papers on level 2.

Hence formula (2) gives the relationship between the h-index of a single paper and its total number of received citations. This is exactly the relationship studied in Schubert (2009). In other words, the h-index of a single paper is a concavely (since $\alpha > 1$) increasing power law function of the total number of citations to this paper.

However, Schubert did not present the power function in a graphical form. He only indicated (last two lines of page 561) that a power function also gives a tolerable fit. Schubert presents a linear fit to the relation between the h-index h of a single paper and the logarithm of the total number T of citations to this paper. Statistically this implies an logarithmic relationship between h and T of the form

$$h = a + b \log T \tag{3}$$

leading to the linear relation between h and log T. Stated otherwise, (3) implies an exponential relationship of T in function of h. But such a relationship is only statistically established and falls outside the classical framework of Lotkaian informetrics.

Based on Schubert's comment (last lines of page 561), both models (2) and (3) are statistically fitted but - as shown here - only (2) explains Schubert's functional relationship between the h-index of a single paper and its total number of received citations, in an informetric way.

References

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

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

L. Egghe (2005). *Power Laws in the Information Production Process: Lotkaian Informetrics.* Elsevier, Oxford, UK.

L. Egghe (2009). The Hirsch-index and related impact measures. *Annual Review in Information Science and Technology (ARIST),* to appear.

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

J.E. Hirsch (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.

A.J. Lotka (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences*, 16(12), 317-324.

A. Schubert (2009). Using the h-index for assessing single publications. *Scientometrics*, 78(3), 559-565.