The influence of the broadness of a query of a topic on its h-index: Models and examples of the h-index of N-grams

DOI: 10.1002/asi.20843
Handle: http://hdl.handle.net/1942/7919

Short Communication

The influence of the broadness of a query of a topic on its h-index: Models and examples of the h-index of N-grams

by

L. Egghe [(1,2,3)]

and

I.K. Ravichandra Rao[(2,4)]

Universiteit Hasselt (UHasselt), Campus Diepenbeek, Agoralaan, B-3590 Diepenbeek, Belgium [(2)]

Universiteit Antwerpen (UA), Stadscampus, Venusstraat 35, B-2000 Antwerpen, Belgium [(3)]

Indian Statistical Institute (ISI), 8th Mile, Mysore Road, R.V. College P.O., Bangalore-560059, India [(4)]

ABSTRACT

The paper studies the influence of the query formulation of a topic on its h-index. In order to generate pure random sets of documents, we used N-grams (N variable) to measure this influence: strings of zeros, truncated at the end were used. The used databases are: Web of

---

[1] Corresponding author.

Key words and phrases: h-index, Hirsch, topic, query, N-gram.

Acknowledgement: The authors are grateful to M. Goovaerts for logistical help during the preparation of this paper.

Acknowledgement: The second named author is grateful to the Universiteit Hasselt for financial support during the time he was a visiting professor.

Acknowledgement: The authors are grateful to Prof. Dr. H. Small (Thomson Scientific) for communicating to us various h-indices of the Web of Science and for allowing us to mention this in this paper.

Science and Scopus. The formula $h = T^{\frac{1}{\alpha}}$, proved in [L. Egghe and R. Rousseau (2006). An informetric model for the Hirsch-index. Scientometrics 69(1), 121-129, 2006], where T is the number of retrieved documents and α is Lotka's exponent, is confirmed, a concavely increasing function of T.

We also give a formula for the relation between h and N the length of the N-gram: $h = D10^{-\frac{N}{\alpha}}$ where D is a constant, a convexly decreasing function, which is found in our experiments.

Nonlinear regression on $h = T^{\frac{1}{\alpha}}$ gives an estimation of α which can then be used to estimate the h-index of the entire database WoS and Scopus: $h = S^{\frac{1}{\alpha}}$ where S is the total number of documents in the database. The value of α was found to be 1.80841 for WoS and 1.84778 for Scopus (using the same searches). Accordingly the values of h for these two databases were estimated but it is also noted that the method is not stable.

I. Introduction

Hirsch (2005) defines the h-index of a scientist as follows: if the scientist's papers are ranked in decreasing order of the number of citations they received then h is the largest rank such that all papers on ranks 1,…,h have at least h citations.

An interesting idea was developed in Banks (2006) where one defines an h-index for topics and compounds (defined in the same way as the h-index for scientists). Bar-Ilan (2007) notices that it is difficult to define the h-index for a topic since this heavily depends on the query formulation of this topic in a database. This problem is of course well-known and is linked with the classical notions "precision" and "recall" in information retrieval (see e.g. Salton and Mc Gill (1987)). A topic cannot be defined (as a query) in a unique way since it also depends on the needs of the person who is requesting or formulating an information retrieval search.

We want to remark, however, once one is agreeing on a certain query, the calculation of the h-index is easy: contrary to an author search (needed for the calculation of the h-index of this author), where one must "filter out" the right author, all documents retrieved via a topical query can be used (see also The STIMULATE6 Group (2007)). In addition, if one searches in the databases WoS or Scopus, each retrieved document is complemented with its citation score (the number of citations it has received) where one can request to put the documents in decreasing order of the number of citations they received. From this it is hence a quick and easy exercise to derive the h-index.

This paper is not involved in query formulation itself but we are interested in the influence of the query formulation of a topic on its h-index. In this, we want to define an abstract topic in diverse ways of wideness. We also want to retrieve documents in a kind of random way over the database. In this way we want to avoid, when going from a narrow topic to a wide topic, that at certain levels, new topics come out, e.g. when truncating TRANSPORT* earlier say at TRANSP* or even TRANS* we add topics like "transparant" or "transcendental" which might disturbe the models. Even TRANSPORT* can refer to very different concepts (e.g. goods transport, osmotic transport, electron transport, …). Of course the topic should also make it possible to have several query formulations (from wide to narrow) yielding a manageable result: avoiding that the system is not giving the exact number of documents anymore (for a wide search) and avoiding searches yielding 0 documents.

We considered using N-grams at different levels (N), followed by a truncation. However, as became clear, N-grams of letters are "drying up" easily meaning that, very quickly, the search results yield no documents anymore. The replacement of letters by numbers turned out to be a solution and the best result was obtained by using the simple N-gram consisting solely of zeros. So we used the queries 000*,…, up to 00000000000* giving us enough datapoints for our model (the queries 0* and 00* were too wide for getting a number of documents, both in WoS and Scopus and also 000* could not be used in WoS – see further). This was not possible with strings of other numbers, e.g. 111* and so on: these were "drying up" too early so that not enough data points were generated.

Supposing the article-citation information process to be Lotkaian (cf. Lotka (1926) or Egghe (2005)) with exponent $\alpha > 1$ and for $n \geq 1$

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

being the number of retrieved documents with n citations, we proved in Egghe and Rousseau (2006) that the h-index h of such a system is given by (in the continuous setting) the power law

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

where T equals the number of documents and where $\alpha > 1$. Note that in (1) the case $n = 0$ is not included (Lotka's law does not include this) but this does not have an influence on the value of h (obviously by definition of the h-index). Hence we expect, as a function of T, h to be concavely increasing (since (2) is a concavely increasing function of T, since $\alpha > 1$). The experiment confirmed this model as we will explain in the next section.

In Section III, since we happen to work with N-grams, we explore the relation of the h-index of the (truncated) N-gram with N. We give a model for this, based on (2), and prove that the model is convexly decreasing, a fact that is confirmed by the retrieval experiment.

Finally, in Section IV, based on the fitted model (2), we use the estimate of the exponent α to give a theoretical model to estimate the h-index of the whole database: if S is its size (exactly known for WoS and approximately for Scopus – see further), then, again by (2) the h-index of the whole database is estimated by $S^{\frac{1}{\alpha}}$, which could be considered as the h-index of "all" papers, produced by the meta-scientist, being the production of "all" scientists together. We, however, also note that this is only a theoretical application since the final result $h = S^{\frac{1}{\alpha}}$ depends heavily on α which is only estimated.

The paper closes with conclusions and some open problems.

II. The h-index of a topic as a function of the number of retrieved documents

Consider a well-defined topic, i.e. a topic defined by a clear command in a certain IR-system. Here we will restrict our attention to two databases namely the Web of Science (WoS) and Scopus since both databases are comprehensive (although not covering the same documents of course) which cannot be said for the database Google Scholar (see Bar-Ilan (2007)).

As explained in the Introduction we will use N-grams consisting of numbers, in our case only using the number 0. This means that we use the topics 0*, 00*, 000*, 0000* (* = truncation) and so on, hereby always further restricting the retrieved sets to subsets of the previous one. We continue until we reach a point where no documents are retrieved.

It is irrelevant for the paper but we checked several papers on why they were retrieved. It turned out that most papers (that we examined) contain a statistical test where significance probabilities where given in the form "$p < 0.000...$".

The searches were performed on May 14, 2007. Citation data are (automatically) produced up to that date. One of the referees requested the reproducibility of the experiments. This is indeed a normal scientific request. However, for this, one needs to limit the citing period to up to May 14, 2007. We do not see how to do this on a monthly or daily basis. It is possible to limit up to 2007 but even this has to be done manually since we cannot restrict to a citing year in an automatic way. This is, in fact, a problem that was also encountered in Liang (2006) (and further discussed in Burrell (2007) and Egghe (2008)) on the construction of (on a yearly basis) h-index sequences of an author. Of course, we are convinced that any search using citation data up to the present date will yield similar results.

We start our investigations in the WoS.

II.1 Hierarchical search in WoS

In WoS we cannot use the commands 0*, 00* and 000* since the system only communicates that “over 100,000 documents are retrieved”. We therefore start with 0000* and it follows

from Table 1 that we can continue until we use 11 digits 0. The first column in Table 1 gives the query, the second gives T, the number of retrieved documents and the third column gives the h-index for this search result.

Table 1. Hierarchical search in WoS using
N-grams consisting of zeros 0.

Command	T	h
00000000000*	3	3
0000000000*	11	9
000000000*	17	11
00000000*	35	15
0000000*	87	22
000000*	380	46
00000*	1,770	80
0000*	9,710	152

The 8 datapoints are depicted in Fig .1, where it is clear that h is concavely increasing in function of T. We have fitted this dataset by nonlinear regression for the power law

$$y = x^{\beta} \tag{3}$$

where $0 < \beta < 1$, basing ourselves to the Lotkaian model (2). Here $y = h$, $x = T$ and $\beta = \frac{1}{\alpha}$ as estimates. Since we want to check formula (2), α is considered here as a constant.

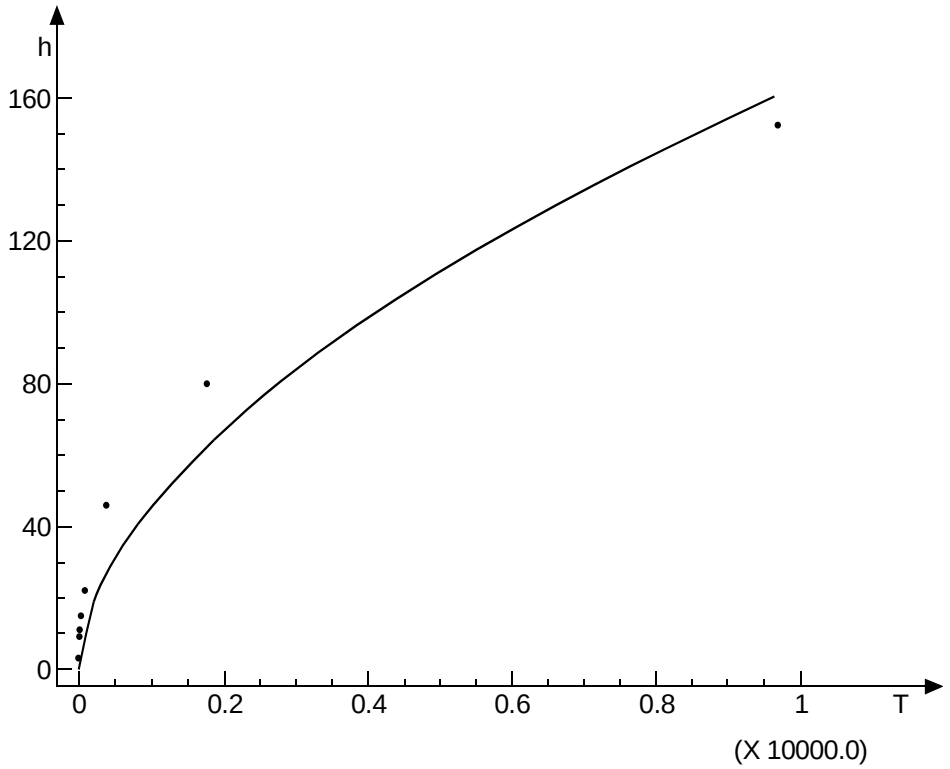

Fig. 1. h in function of T for the data in Table 1 and fitted model.

The system gives a best fit for $\beta = \frac{1}{\alpha} = 0.5529719$, hence for $\alpha = 1.80841$ (not far away from the "classical" $\alpha = 2$, cf. Lotka (1926), Egghe (2005)).

Normally, such experimental data are explained using models such as in (2). Here we had (2) before the experimental data were retrieved and now the data (Fig. 1) confirm the theoretical Lotkaian model !

We consider the result that α is close to 2 as logical, considered that we are working in such a large, multidisciplinary database. This result will be refound in the sequel.

II.2 Hierarchical search in Scopus

In Scopus we can use the N-grams from 000* onwards as Table 2 shows.

Table 2. Hierarchical search in Scopus using
N-grams consisting of zeros 0.

Command	T	h
00000000000*	3	2
0000000000*	9	6
000000000*	20	10
00000000*	48	14
0000000*	126	24
000000*	527	48
00000*	2,479	82
0000*	13,656	166
000*	526,539	462

We present two fits of these data: one excluding 000* (for better comparison with the WoS results) and then one including 000*.

Excluding 000* we have the graph of the data and the fit of (3) using nonlinear regression, as in Fig. 2.

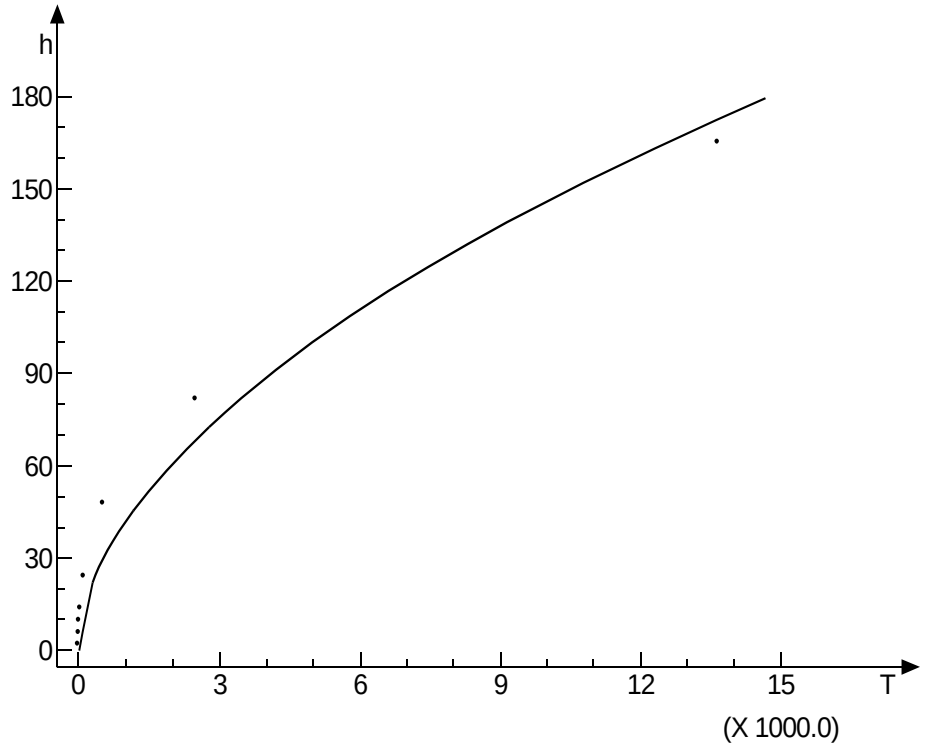

Fig. 2. h in function of T for the data in Table 2 (excluding 000*) and fitted model.

The system gives a best fit for $\beta = 0.54119$, hence for $\alpha = 1.84778$, very close to the WoS results and again not far away from the classical $\alpha = 2$.

Now we repeat the same exercise in Scopus for all data in Table 2 (hence including 000*). See Fig. 3 for the data points and the fit of (3)

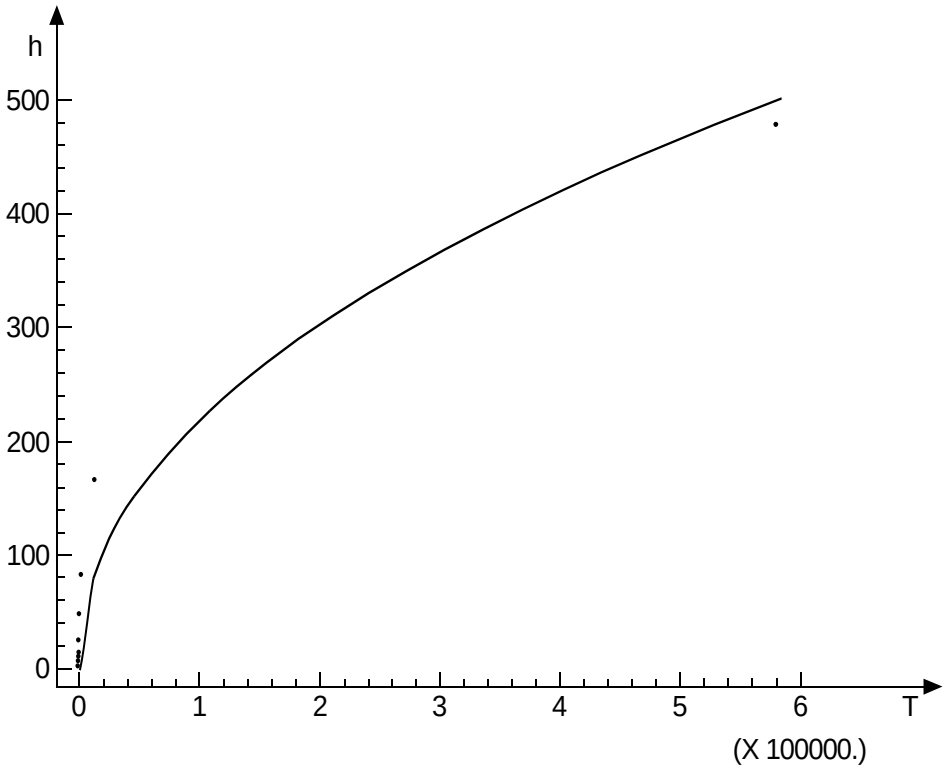

Fig. 3. h in function of T for the data in Table 2 (including 000*) and fitted model.

Nine data points instead of eigth gives a slight difference for the fitted model: now $\beta = 0.4679173$, hence $\alpha = 2.13713$ but we see that we still have that $\alpha \approx 2$.

Note: As some of the referees remarked, the fits in Figs. 1-3 are not impressive (although the concave increase is clear from the datapoints). It is clear that functions with more than one parameter (e.g. $aT^{\frac{1}{\alpha}}$ instead of (2) as suggested by one of the referees) will provide a better fit but that was not the purpose of these experiments: here we wanted to determine α and (2) which is a theoretically proved h-index relation in function of the number of publications T (Egghe and Rousseau (2006)).

Now we turn our attention to the relation of h with N, the length of the N-gram.

III. The h-index of a N-gram as function of N.

III.1 Theoretical model

Each digit in an N-gram of numbers is a choice amongst 10 digits. So, on the average, each digit divides the database in 10 (assumed equal) parts of which one agrees with this digit.

More concretely, the 1-gram 0* retrieves, approximately 10% (i.e. $\frac{1}{10}$th) of all documents (or at least of all documents that contain a number which we assume to be all documents but this is not necessary for the model). Then, from this set of documents, the 2-gram 00* retrieves, approximately, 10% of the documents, hence a fraction $\frac{1}{10^2}$ of the original document set (database).

Let the original database contain S documents. Then an N-gram of numbers retrieves, on average

$$T_N = \frac{S}{10^N} \tag{4}$$

documents. The h-index of this set is, according to (2):

$$h = h(N) = T_N^{\frac{1}{\alpha}} = \left(\frac{S}{10^N}\right)^{\frac{1}{\alpha}} = \frac{D}{10^{\frac{N}{\alpha}}} \tag{5}$$

It is readily seen that (5) is a convexly decreasing exponential function of N. Now we will check this model on our data in tables 1 and 2.

Note: As is clear from the data in Tables 1 and 2, formula (4) is only a simple estimation of reality (the simplest one possible, but, in a theoretical way, the most natural one). In addition, formula (5) is fitted very well as will be seen in the next subsection.

III.2 Application to the data of Tables 1 and 2

Let us start with the WoS data (Table 1). From this Table it is clear that we have the data points as in Table 3.

Table 3. WoS data: h versus N

N	h
4	152
5	80
6	46
7	22
8	15
9	11
10	9
11	3

The graph is depicted in Fig. 4, where also the fit of (5) is shown.

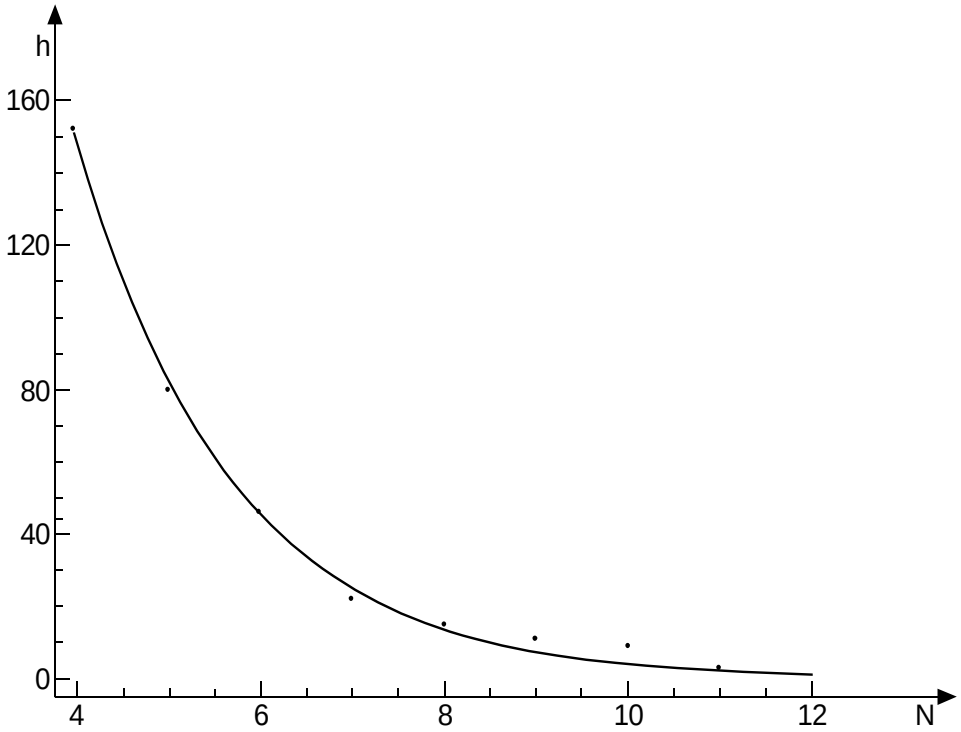

Fig. 4 h in function of N for the data in Table 3 and fitted model.

The fitted model is $h(N) = 1{,}678.73 \times 10^{-0.261625N}$ confirming (5) and the convex decrease.

Now we go to the Scopus data. From Table 2 we derive the data shown in Table 4.

Table 4. Scopus data: h versus N

N	h
3	462
4	166
5	82
6	48
7	24
8	14
9	10
10	6
11	2

These data and fit of (5) are depicted in Fig. 5.

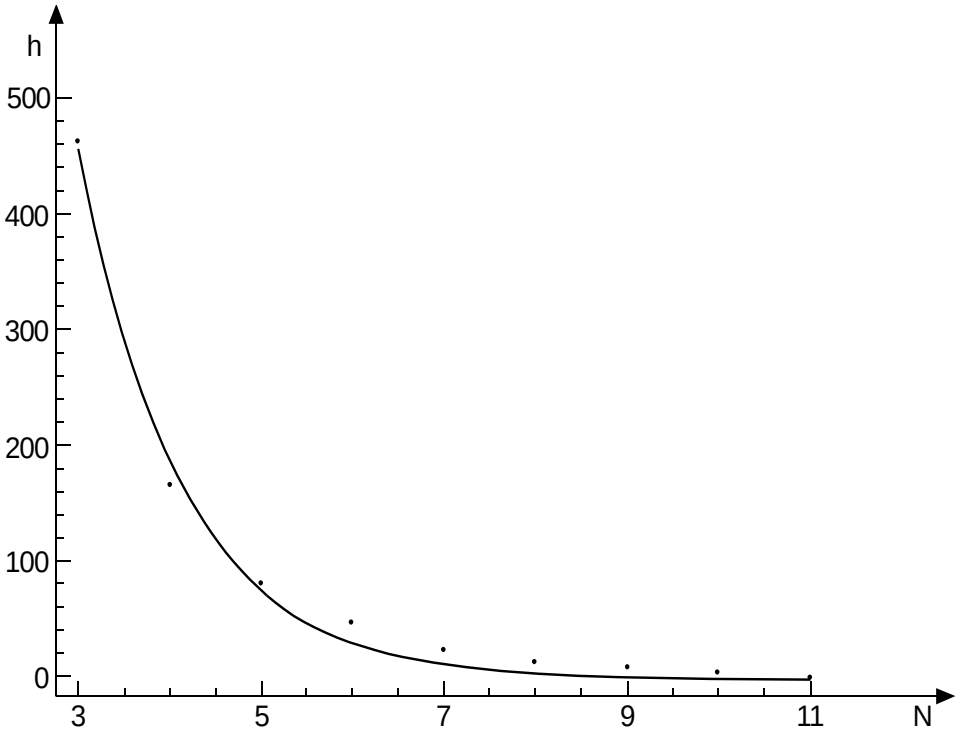

Fig. 5 h in function of N for the data in Table 4 and fitted model.

The fitted model is $h(N) = 6{,}852.33 \times 10^{-0.391995N}$.

IV. The h-index of databases

Based on the above results it is, at least theoretically, possible to determine the h-index of the entire database WoS or Scopus, i.e. using all documents in these databases.

Indeed, let S denote the number of documents in the database. Once we know α we can apply (2) to the entire database yielding

$$h = S^{\frac{1}{\alpha}} \tag{6}$$

For the WoS we have $S = 33,061,919$ (source and non-source) documents at the time of our experiments (May 14, 2007). This number was obtained from information given on the site of WoS on that day. Using $\alpha = 1.80841$ (Section II) we estimate

$$h_{WoS} = 14,387$$

For Scopus there is only an estimate of $S = 30,000,000$ documents given. Using that $\alpha = 1.84778$ we get

$$h_{Scopus} = 11,131$$

Of course, if we include 000* we found $\alpha = 2.13713$ and then we have

$$h_{Scopus} = 3,152$$

showing that the method is not stable.

In general we can assume that $\alpha \gg 2$ for a large multidisciplinary database which gives the rough estimate that its h-index is

$$h_{database} \gg \sqrt{S} \tag{7}$$

Note 1: H. Small (Small (2007)) communicated to us that – dependent on the used time period (up to 1955 or up to 1972) and dependent on the document set (only containing source documents or containing source and non-source documents), the h-index of WoS varies between 1,500 and 2,000. With this information, we can use (6) the other way around: knowing $S = 33{,}061{,}919$ (source documents only up to 1972) and knowing that $h_{WoS} = 1{,}555$ for this collection (communication of H. Small), we arrive at an estimate of α:

$$\alpha = \frac{\ln S}{\ln h_{WoS}} \tag{8}$$

$$\alpha = 2.3558782$$

being Lotka's α for WoS, considered as a Lotkaian document-citations system.

This value is slightly higher than the ones conjectured in this paper and above the critical $\alpha = 2$.

Note 2: Since the model (6) is based on Lotka' law, where sources have at least 1 item, we should correct S to be the number of ever cited documents in the database. A communication of H. Small gives an estimate that about 80% of the documents in WoS have at least one citation. Replacing S above by 0.8S we now have, using $\alpha = 1.80841$,

$$h_{WoS} = 12{,}717$$

Doing the same for Scopus we find

$$h_{Scopus} = 9{,}865$$

for $\alpha = 1.84778$ and for $\alpha = 2.13713$ as above:

$$h_{Scopus} = 2{,}840$$

which is more close to the 1,500-2,000 range given by H. Small (but the difference remains).

The estimate of α, as in (8) is now (replace S by 0.8S):

$$\alpha = 2.3255154$$

which is almost the same as the value obtained above.

V. Conclusions and open problems

In this paper we calculated h-indices for topics with a variable "wideness" and we showed that the model

$$h = h(T) = T^{\frac{1}{\alpha}} \qquad (8)$$

applies well (concave increase) where T is the number of retrieved documents and α is the Lotka exponent.

The variable wideness of the topic is reached using N-grams of numbers: 000*, 0000*, 00000* and so on. The above model (8) has been extended to a model for the h-index in function of the length N of the N-gram:

$$h = h(N) = D.10^{-\frac{N}{\alpha}} \qquad (9)$$

where D is a constant and α is as above. The obtained data confirm this convex decreasing relation.

Based on the obtained Lotka exponent α and the size S of a database we can then (at least theoretically) determine the h-index of the entire database. We have that $h = S^{\frac{1}{\alpha}}$. We conjecture that, for large multidisciplinary databases, $\alpha \approx 2$ so that $h \approx \sqrt{S}$ although data of H. Small suggest that $\alpha \approx 2.3$.

We leave it open to find more stable methods to estimate the h-index of a database. We also encourage colleagues to further experiment on the h-index of topics expressed via different widenesses of the query) and especially on the h-index of N-grams.

References

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

J. Bar-Ilan (2007). The h-index of h-index and of other informetric topics. Proceedings of ISSI 2007: 11th International Conference of the International Society for Scientometrics and Informetrics (CSIC, Madrid, Spain), D. Torres-Salinas and H.F. Moed, eds., 826-827, 2007.

Q.L. Burrell (2007). Hirsch index or Hirsch rate ? Some thoughts arising from Liang's data. Scientometrics 73(1), 19-28.

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

L. Egghe (2008). Mathematical study of h-index sequences. 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, 16569-16572.

L. Liang (2006). h-index sequence and h-index matrix: Constructions and applications. Scientometrics 69(1), 153-159.

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

G. Salton and M.J. Mc Gill (1987). Introduction to modern Information Retrieval. Mc Graw-Hill, Singapore.

H. Small (2007). Written communication.

The STIMULATE6 Group (2007). The Hirsch index applied to topics of interest to developing countries. First Monday 12(2), 2007. http://www.firstmonday.org/issues/issue12 2/stimulate/