Why h-Index

Vladik Kreinovich, Olga Kosheleva, and Hoang Phuong Nguyen

Abstract At present, one of the main ways to gauge the quality of a researcher is to use his or her *h-index*, which is defined as the largest integer n such that the researcher has at least n publications each of which has at least n citations. The fact that this quantity is widely used indicates that h-index indeed reasonably adequately describes the researcher's quality. So, this notion must capture some intuitive idea. However, the above definition is not intuitive at all, it sound like a somewhat convoluted mathematical exercise. So why is h-index so efficient? In this paper, we use known mathematical facts about h-index – in particular, the results of its fuzzy-related analysis – to come up with an intuitive explanation for the h-index's efficiency.

1 Formulation of the Problem

h-index is ubiquious. At present, one of the main criteria used to evaluate the quality of a researcher is an *h-index*, a concept first proposed in [5] and then spread like wildfire. We will explain what it is in the next paragraph, but we just want to mention that while everyone seems to agree that it is an imperfect characteristic of a person's research status, the h-index is what is cited in recommendation letters for promotion or for hiring, this is what is cited when nominating someone for awards, this is what is cited everywhere.

What is h-index? The definition of an h-index is somewhat convoluted, actually so convoluted that people who are nor familiar with this notion – e.g., new students –

Vladik Kreinovich and Olga Kosheleva
University of Texas at El Paso, El Paso, Texas 79968, USA
e-mail: vladik@utep.edu, olgak@utep.edu

Hoang Phuong Nguyen
Division Informatics, Math-Informatics Faculty, Thang Long University, Nghiem Xuan Yem Road
Hoang Mai District, Hanoi, Vietnam, e-mail: nhphuong2008@gmail.com

get surprised to learn that such a non-intuitive notion is so ubiquitous. Specifically, an h-index is defined as *the largest integer n such that a researcher has at least n publications each of which has at least n citations*.

Comment. It should be taken into account that the numerical value of an h-index depends on how we count citations. One possibility is to use only Web of Science citations – in which case we miss many citations in conference papers, and in computer science, many important results are published in peer-refereed archival conference proceedings. Another possibility is to use Google citations – in which case conference papers counts, but so are citations in less reputable journals.

In both cases, however, the same general definition is used.

But why h-index? One could think of many ways to gauge a researcher's productivity and its effect. Why namely h-index – whose definition does not seem to be intuitive at all – has been so widely spread? The very fact that this index is widely used is an indication that it does reflect the researcher's quality. So, a natural question is: why is this seemingly non-intuitive definition so efficient?

What we do in this paper: general idea. In this paper, we provide an intuitive explanation for h-index.

People (kind of) tried to answer this question. The notion of h-index has been actively studied from the mathematical and computational viewpoint. In particular, fuzzy researchers may be aware that almost immediately after h-index has been invented, a paper [13] showed that h-index is a particular case of Sugeno integral – a notion closely related to fuzzy logic (see, e.g., [2, 6, 8, 9, 10, 11, 14]) and thus, very familiar in fuzzy community. Sugeno integral was first introduced in [12]; for a latest overview, see, e.g., [1].

Does this explain the ubiquity of h-index? Maybe for a few die-hard fans of Sugeno integral it does, but not for others – Sugeno integral is just one of the many different "fuzzy integrals", different ways to combine different estimates into a single value. Even the paper [13] mentions that we could use other fuzzy integrals – and get other bibliometric characteristics. So the question remains: why h-index?

What we do in this paper: some more details. In contrast to h-index, the Sugeno integral *does* have an intuitive understanding. So what we did is borrow this understanding and transform it into an understanding of why h-index is so ubiquitous and so efficient.

2 Our Explanation

Analysis of the problem: towards an intuitive understanding of what is a influential researcher. The researcher's output is usually his or her publications. Each publication contains some results. Some of these results are influential, some are not that influential. From the commonsense viewpoint, an influential researcher is

a one who produces many influential results, i.e., in other words, *a researcher who published a large number of influential papers*.

How can we gauge whether an idea described in a paper is influential? By definition, an idea is influential if it *influences* others and leads to other ideas – i.e., to new papers that use the original idea – and thus, that cite the original paper. So, a natural way to gauge how influential is a given paper is to consider how many other papers cite it. If a paper has a large number of citations, this means that this paper is influential.

Substituting this (informal) definition of "influential paper" into the above (informal) definition of an influential researcher, we arrive at the following informal definition: an influential researcher is *a researcher who published a large number of papers each of which has a large number of citations*.

How to formalize the above intuitive (informal) definition? The wording "a large number of" is informal, it means different things to different people, and a proper formal description of this logic would indeed require the use of fuzzy logic (or some other techniques appropriate for describing informal notions).

The simplest formalization of this notion. Intuitively, the notion "the large number of" is imprecise. So, to get an adequate formalization of this notion, we should use techniques for formalizing such imprecise terms, such a fuzzy logic. However, for simplicity, let us see what happens if we use the simplest possible formalization of this notion: namely, we select some threshold value n_0 and then:

- if we have $n \geq n_0$ of items, we say that we have a large number of items, and
- if we have $n_0 < n$ items, then we say that do not have a large number of items.

With respect to formalizing the notion of "an influential paper" – which we interpreted as "a paper with large number of citations", we thus get the following formalization – which we will call n_0-*influential*: a paper is n_0-influential if it has at least n_0 citations.

The resulting formalization of the notion of an influential researcher. The notion of n_0-large number leads to the following formalization of the above intuitive notion of an influential researcher: a researcher is n_0-*influential* if this researcher has published at least n_0 papers each of which has at least n_0 citations.

Analysis of this notion leads to the desired explanation. If we have two possible thresholds $n_0 > n_0'$, then clearly each n_0-influential researcher is also n_0'-influential. Thus, to properly gauge the quality of a researcher, it makes sense to consider the largest possible value n_0 for which this researcher is n_0-influential.

This largest number is *the largest number n_0 such that the researcher has published at least n_0 papers each of which has at least n_0 citations*. This is exactly the definition of the h-index. So, we have indeed come up with an intuitive explanation of the h-index.

3 What Next?

There are many ways in which the above explanation can be used to improve the notion of the h-index. Let us list the two main ideas.

A first natural idea is to replace a simplified formalization of the notion "a large number of" with a more adequate fuzzy notion, in which, for each n, we have a degree to which this n is large – i.e., to which extend this number n corresponds to the above notion. This will hopefully allow us to distinguish between the cases between which h-index does not distinguish: e.g., between the two researchers each of which has exactly one published paper but whose papers have different number of citations: 100 for the first paper, 1 for the second one. For both researchers, the h-index is the same – equal to 1, but clearly the first researcher is more influential. This idea may lead to some of the fuzzy modifications proposed in [13] or to yet another characteristics – depending on what "and"- and "or"-operations we use.

Another natural idea is to take into account that not all citations are equal: a citation by an influential paper (which is itself highly cited) should be valued more than a citation by a paper which was not cited at all. There should be a weight, e.g., proportional to the number of citations of the citing paper. These citations should also be similarly weighted – as a result, we end up with the notion of an eigenvalue, similar to Google's PageRank (see, e.g., [3, 4, 7]) or to the eigenvalues used to estimate the quality of a journal. So, the second idea is to replace the number of citations with eigenvalue in the definition of an h-index.

Acknowledgments

This work was supported in part by the National Science Foundation via grants 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science) and HRD-1242122 (Cyber-ShARE Center of Excellence).

References

1. G. Beliakov, S. James, and J.-Z. Wu, *Discrete Fuzzy Measures: Computational Aspects*, Springer, Cham, Switzerland, 2020
2. R. Belohlavek, J. W. Dauben, and G. J. Klir, *Fuzzy Logic and Mathematics: A Historical Perspective*, Oxford University Press, New York, 2017.
3. S. Brin and L. Page, "The anatomy of a large-scale hypertextual Web search engine". *Computer Networks and ISDN Systems*, 1998, Vol. 30, No. 1–7, pp. 107–117.
4. D. F. Gleich, "PageRank beyond the Web", *SIAM Review*, 2015, Vol. 57, No. 3, pp. 321–363.
5. J. E. Hirsch, "An index to quantity an individual's research output", *Proceedings of the National Academy of Science of the USA*, 2005, Vol. 102, No. 45, pp. 16569–16572.
6. G. Klir and B. Yuan, *Fuzzy Sets and Fuzzy Logic*, Prentice Hall, Upper Saddle River, New Jersey, 1995.

7. A. N. Langville and C. D. Meyer, *Google's PageRank and Beyond: The Science of Search Engine Rankings*, Princeton University Press, Princeton, New Jersey, 2006.
8. J. M. Mendel, *Uncertain Rule-Based Fuzzy Systems: Introduction and New Directions*, Springer, Cham, Switzerland, 2017.
9. H. T. Nguyen and V. Kreinovich, "Nested intervals and sets: concepts, relations to fuzzy sets, and applications", In: R. B. Kearfott and V. Kreinovich (eds.), *Applications of Interval Computations*, Kluwer, Dordrecht, 1996, pp. 245–290.
10. H. T. Nguyen, C. Walker, and E. A. Walker, *A First Course in Fuzzy Logic*, Chapman and Hall/CRC, Boca Raton, Florida, 2019.
11. V. Novák, I. Perfilieva, and J. Močkoř, *Mathematical Principles of Fuzzy Logic*, Kluwer, Boston, Dordrecht, 1999.
12. M. Sugeno, *Theory of Fuzzy Integrals and Its Applications*, PhD Dissertation, Tokyo Institute of Technology, Tokyo, Japan, 1974.
13. V. Torra and Y. Narukawa, "The h-index and the number of citations: two fuzzy integrals", *IEEE Transactions on Fuzzy Systems*, 2008, Vol. 16, No. 3, pp. 795–797.
14. L. A. Zadeh, "Fuzzy sets", *Information and Control*, 1965, Vol. 8, pp. 338–353.