A Bibliometric Index Based on Collaboration Distances

Maria Bras-Amorós[1], Josep Domingo-Ferrer[1], and Vicenç Torra[2]

[1] Universitat Rovira i Virgili
UNESCO Chair in Data Privacy
Department of Computer Engineering and Mathematics
Av. Països Catalans 26, E-43007 Tarragona, Catalonia
{maria.bras,josep.domingo}@urv.cat
[2] IIIA, Institut d'Investigació en Intel·ligència Artificial
CSIC, Consejo Superior de Investigaciones Científicas,
Campus UAB s/n, E-08193 Bellaterra, Catalonia
vtorra@iiia.csic.es

The h-index by Hirsch[1] has recently earned a lot of popularity in bibliometrics, being echoed in Nature and implemented in the Web of Science bibliometric database. Previous indicators were the total number of papers or the total number of citations. Following the widely accepted idea that not all papers should count equally, the h-index counts only those papers that are significant enough according to their number of citations. However, as for qualifying the significance of citations, beyond excluding self-citations by recent proposals[2,3,4,5,6], the fact that not all citations should count equally has remained unaddressed, with the exception of [7]. The h-index can be described in terms of a pool of evaluated objects (papers), a quality function on the evaluated object (citations received by each paper) and a sentencing line crossing the origin ($y = x$). When the evaluated objects are ordered by descreasing quality, then the intersection of the sentencing line with the graph of the quality function yields the index value.

Based on this abstraction, we present a new index, the c-index, in which the evaluated objects are the citations received (by a paper, an author, a research group, a journal, etc.), the quality of a citation is the collaboration distance between the authors of the cited and the citing papers when the citation appears, and the sentencing line takes a slope α between 0 and ∞. To mitigate the small world effect we suggest taking $\alpha \approx 1/4$. As a result, the new index counts only those citations which are significant enough, where significance is proportional to the collaboration distance between the cited and the citing authors.

While an h-index x means that there are x papers with at least x citations each and the rest of papers with at most x citations, a c-index x means that there are x citations (regardless of the papers to which these citations refer) at collaboration distance at least $\alpha\ x$ and the rest of citations at collaboration distance at most $\alpha\ x$.

If we want to differentiate between recurrent collaborations and occasional collaborations, a refined version of the classical distance can be defined, where the distance between two coauthors is inversely proportional to the number of joint papers between them.

V. Torra, Y. Narukawa, and M. Daumas (Eds.): MDAI 2010, LNAI 6408, pp. 5–6, 2010.

Some of the advantages of the new c-index are:

1. It gives a solution to the problem of few but seminal contributions, which for instance means that Galois has h-index 2, and which is also especially important when evaluating journals[6].
2. It neutralizes self-citations and citations by close authors.
3. It discourages gratuitous coauthorship.
4. The new index is not linear anymore with respect to the scientific age, rewarding citations to and from novel authors and thus modernity.
5. Multiple spelling of one single reference in different citations or misspelling reference data other than authorship, which decrease the h-index, do not affect the c-index.

Together with the f-index [7], the c-index is a pioneer in the bibliometric literature in measuring the output of a scientist or a journal based at the same time on the quantity and quality of the received citations: the more distant the citing authors, the higher the quality of a citation (this notion of quality rewards contributions of broad interest).

Since any bibliometric index is referred to a particular database, it should be easy for any of the bibliometric databases to automate the computation of the c-index, just like some of them have automated the computation of classical distances (MathSciNet of the American Mathematical Society) or the computation of the h-index (Web of Science).

Being based only on citations, the c-index loses the feature of the h-index of counting how many papers among those by an author have had a decent impact. To remedy this, one might combine the h-index and the c-index by providing both of them or by mixing them (e.g. as $\sqrt{hc}$) if required for ranking purposes.

In [8] one can find an extended version of this paper with a deeper discussion on the c-index, a detailed comparison with the most recent indices, some computational hints, and some experiments.

References

1. Ball, P.: Index aims for fair ranking of scientists. Nature 436, 900 (2005)
2. Schreiber, M.: Self-citation corrections for the Hirsch index. Europhysics Letters (EPL) 78, 30002p1–30002p6 (2007)
3. Derby, B.: H-factors research metrics and self-citation. Nature Blogs (April 25, 2008)
4. Zhivotovsky, L.A., Krutovsky, K.V.: Self-citation can inflate h-index. Scientometrics 77(2), 373–375 (2008)
5. Egghe, L.: An improvement of the h-index: the g-index. ISSI Newsletter 2(1), 8–9 (2006)
6. Braun, T., Glänzel, W., Schubert, A.: A Hirsch-type index for journals. Scientometrics 69(1), 169–173 (2006)
7. Katsaros, D., Akritidis, L., Bozanis, P.: The f-index: quantifying the impact of coterminal citations on scientists' ranking. Journal of the American Society for Information Science and Technology 60(5), 1051–1056 (2009)
8. Bras-Amorós, M., Domingo-Ferrer, J., Torra, V.: A bibliometric index based on the collaboration distance between cited and citing authors (submitted 2010)