ISSN 0147-6882, Scientific and Technical Information Processing, 2013, Vol. 40, No. 3, pp. 142–145. © Allerton Press, Inc., 2013.
Original Russian Text © S.D. Shtovba, E.V. Shtovba, 2013, published in Nauchno-Technicheskaya Informatsiya, Seriya 1, 2013, No. 7, pp. 28–31.

A Citation Index with Allowance for the Implicit Diffusion of Scientific Knowledge

S. D. Shtovba and E. V. Shtovba
Vinnitsa National Technical University, Ukraine
e-mails: shtovba@gmail.com; olena.shtovba@yahoo.com
Received April 2, 2013

Abstract—This paper suggests a new scientometric index that estimates knowledge diffusion and has two constituents: the first one is equivalent to a usual citation index, i.e., it describes the visible diffusion of scientific knowledge; the second one reflects the implicit diffusion of scientific knowledge and is expressed through the number of implicit citations. The practical value of the suggested index is that it permits implicit initiators of the scientific mainstream to be easily identified. The distinctive feature of such scientists is the large value of the suggested citation index and the low value of the usual citation index.

Keywords: scientometry, citation index, implicit citation, knowledge diffusion, initiator of the mainstream
DOI: 10.3103/S0147688213030040

INTRODUCTION

Scientometric indices are used more and more often today jointly with expert reports in order to assess the value of scientific activity. Above all, scientometric indices are claimed to assess fundamental research, whose results are not directly related to an economic effect. Fundamental research is aimed at the development of science, therefore, the demand for it is assessed through the commentary of the scientific community on publications with the results of studies. This commentary is formally expressed by the citation index, viz., the total number of publications that cite the works under consideration. Variants of the citation index modify it as follows:

—taking the personal contribution of a scientist into consideration [1, 2], by dividing the number of citations between co-authors, whose number in one article sometimes surpasses 3000;

—making allowance for the authority of a citing edition by using the impact factor of a journal or another analogous coefficient [3, 4];

—ignoring self-citation or citation by co-authors [5, 6], which significantly decreases the rating of a "hermit scientist" whose publications are of interest only to him himself;

—ignoring the repeated citation of one work by the same scientist [7], which lowers the influence of complimentary citation.

Using these and other modifications does not remove the following two disadvantages of the citation index. The first disadvantage is due to forgetting the names of classic authorities, when authors believe that the contribution of predecessors is so well known to anyone from the corresponding field of science that it makes no sense to mention it [8]. Correspondingly, the works of such first-rank scientists rapidly cease to be cited. As usual, this process hastens the inclusion of the scientific results of classic authorities in manuals and teaching aids. Therefore, the diffusion of this scientific knowledge, i.e., its adaptation and use in a wide range of scientific and engineering research and developments, is impetuously growing [9].

The second disadvantage of the citation index is the low ranking of conceptual works with principally new ideas. This takes place due to concealing primary sources fairly often, i.e., due to not including the conceptual works of predecessors in a list of cited literature. The creation of a new scientific result is a process that includes the following stages [8]: (1) the generation of an idea and development of a concept; (2) the modification of the idea (concept) with the purpose of improving its accuracy, speed, simplification, etc; (3) advancement and practical application. The history of science contains many examples where the works of the second and third stages are cited much more often than conceptual articles. The articles of A. Folin [10] and O. Lowry [11] are an impressive example from among those presented in [8]. O. Lowry modified A. Folin's reagent for the purpose of colorimetric protein detection by adding one more component, which expanded the application scope of the method. Today, according to the data of Google Scholar, the article by O. Lowry has been cited 247 480 times, meanwhile the number of citations to the work by A. Folin is only 1950. Consequently, among the 247 480 works that cite the article [11] and, correspondingly, use the ideas of

O. Folin on which it is based, few manifestly cite his article [10].

The goal of the present article is to modify the citation index by developing mechanisms for making up for these disadvantages. The idea is to supplement the citation index by a component that takes knowledge diffusion into account. Citation to a work is chosen as an indicator of knowledge diffusion in [12]. However, this is likely to be the direct use of knowledge. There is also implicit knowledge diffusion, which consists of using ideas without directly citing their author but with the possibility of identifying a primary source through a citation chain. In the above-presented example, this is the citation of only the work by A. Lowry without directly citing the article by O. Folin. Let such cases be called implicit knowledge diffusion and a citation chain that leads to a primary source be called implicit citation. Correspondingly, by counting in some way the number of implicit citations and adding it with some weight to the usual citation index, we obtain the ranking of a scientist with allowance for implicit knowledge diffusion. The analogous personnel assessment principles are used in network marketing, when a worker draws points for both his direct sales and sales made by the team that he recruited. Here, the personal sales of a worker and sales of his team correspond to direct citation and implicit citation, respectively. The same principles are used in sports to assess the attacking players of a team according to the "goal plus pass" system. Correspondingly, direct citation and implicit citation are analogues of a goal and a pass, respectively.

FORMALIZING THE PROBLEM STATEMENT

Let us consider known the list of M publications in the following:

$$\mathbf{A}_j = \langle j, \mathbf{F}_j, \mathbf{T}_j \rangle, \; j = \overline{1, M}, \tag{1}$$

where $\mathbf{F}_j = \{F_j^1, F_j^2, F_j^3, \ldots\}$ is the set of the numbers of publications cited by a work $\mathbf{A}_j$;

$\mathbf{T}_j = \{T_j^1, T_j^2, T_j^3, \ldots\}$ is the set of the numbers of publications citing the work $\mathbf{A}_j$.

From the mathematical standpoint, the task is to find a function $\mathbf{A}_j \longrightarrow C_j(j = \overline{1, M})$, where C_j is a new citation index that takes implicit knowledge diffusion into account.

TAKING IMPLICIT KNOWLEDGE DIFFUSION INTO ACCOUNT

Let the citation index with allowance for the implicit knowledge diffusion be defined as follows:

$$C_j = D_j + \alpha I_j, \tag{2}$$

where $D_j = |\mathbf{T}_j|$ is the usual citation index that makes allowance for only direct citations (| | designates the power of a set, i.e., the number of its elements);

I_j is the implicit citation index;

$\alpha \in [0,1]$ is the weight coefficient of the value of implicit citation.

If $\alpha = 0$, the new index (2) is equivalent to the usual citation index. When $\alpha = 1$, the contributions of direct and implicit citation will be equally important.

Let knowledge diffusion be taken into account as follows. Each scientific work contains both new knowledge and knowledge from cited works. Let us suppose that the share of borrowed knowledge is constant. Let it be designated through $\beta \in (0,1)$. Let us consider a simple citation chain $C \longrightarrow B \longrightarrow A$, when a work A only cites the work B and the work B only cites the work C. Correspondingly, the [beta]-share of knowledge passed from B into C; the [beta]-share of knowledge then passed from B into A. Then, the knowledge, whose share is $\alpha = \beta \cdot \beta$ passed from C into A, although this is not proven by direct citation. Such an implicit use of knowledge or implicit diffusion occurs when an author includes the most relevant sources in a reference list, whose level of use surpasses some threshold [8].

Let us consider the case where a reference list in B consists of N sources, which include the work C. Then we will consider that the knowledge, whose share is $\frac{\beta}{N}$ passed from C into B. Generalizing the computations, we obtain the following formula of the implicit citation index:

$$I_j = \sum_{\substack{\forall i \in \mathbf{T}_j \\ j \notin \mathbf{F}_i}} \frac{|\mathbf{T}_i|}{N_i}, \tag{3}$$

where $N_i = |\mathbf{F}_i|$ is the length of the reference list in the ith publication.

The condition $j \notin \mathbf{F}_i$ in (3) indicates that implicit citation is not taken into account in the case of direct citation.

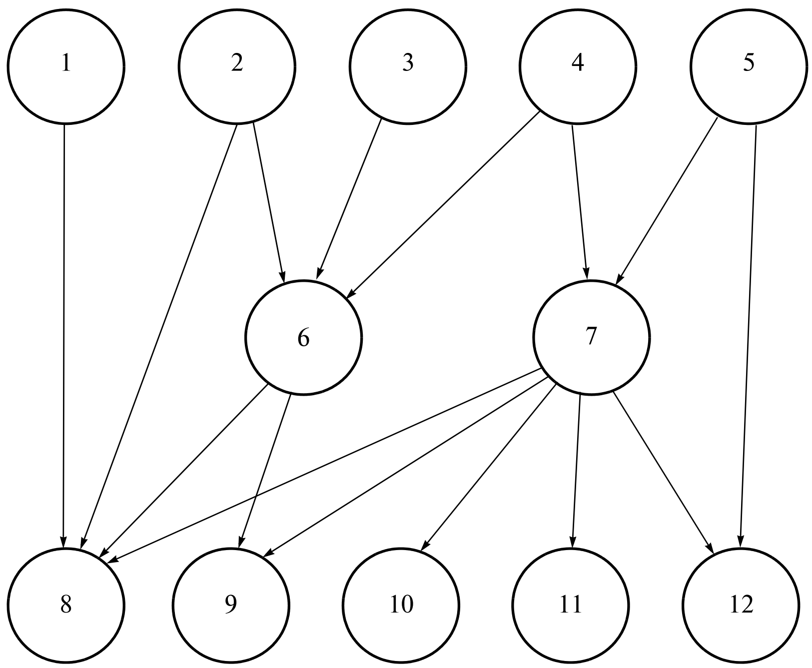
An example. The initial data are given in Table 1. The corresponding citation network is shown in the figure. Table 2 illustrates how the new citation index (2) is calculated according to these data. It shows that publications nos. 2, 4, 5, and 6 have an equal number of direct citations. Among them, the highest ranking is possessed by publication no. 4 owing to the maximum number of implicit citations, two of which were made through work no. 6, and five were made through work no. 7.

COMPARISON WITH ANALOGUES

We found that implicit citation for the purpose of ranking scientific publications was only used in [13] to

Table 1. Initial data on citation in format (1)

j	$\mathbf{F}_j$	$\mathbf{T}_j$
1	Unknown	{8}
2	Unknown	{6, 8}
3	Unknown	{6}
4	Unknown	{6, 7}
5	Unknown	{7, 12}
6	{2, 3, 4}	{8, 9}
7	{4, 5}	{8, 9, 10, 11, 12}
8	{1, 2, 6, 7}	–
9	{6, 7}	–
10	{7}	–
11	{7}	–
12	{5, 7}	–



Citation network

calculate the Hirsch index of a separate article. In this case, the Hirsch index of an article is h, if it is cited by h works, each of which has been cited at least h times.

The nearest analogue is the approach in [14], in which implicit citations are used to estimate a cumulative patent citation index. In this regard, the longer the chain of an implicit citation is, the lower its weight coefficient is. The principal distinction of our approach is that the length of a reference list is taken into account when estimating implicit citations in (3), meanwhile the length of a corresponding patent list is not taken into account in [14]. Ignoring the length of a reference list can lead to case where the constituent of the implicit citation is greater than if direct citation took precedence. For example, for the fragment of the citation network with tops 4, 6, 7, and 8 (see Fig. 1), work no. 8 would provide two points to work no. 4 for implicit citation; meanwhile, in the case of direct citation $4 \longrightarrow 8$, work no. 4 would only get one point.

CONCLUSIONS

A new scientometric index that takes knowledge diffusion into account has been suggested. It has two constituents, the first of which is equivalent to the usual citation index, i.e., it describes the visible diffusion of scientific knowledge. The second constituent reflects the implicit diffusion of scientific knowledge

Table 2. On the calculation of the citation index (2)

j	1	2	3	4	5	6	7
D_j	1	2	1	2	2	2	5
I_j	0	$\frac{1}{3} = 0.33$	$\frac{2}{3} = 0.67$	$\frac{2}{3} + \frac{5}{2} = 3.17$	$\frac{4}{2} = 2$	0	0
C_j, $\alpha = 0.1$	1	2.03	1.07	2.32	2.2	2	5
C_j, $\alpha = 0.5$	1	2.17	1.33	3.58	3	2	5

and is expressed through the number of implicit citations. The suggested approach inherits the mechanisms for assessing the activity of personnel in network marketing and attacking players in hockey, football, and other team sports.

The practical value of the suggested index is that it makes it possible to easily identify creative scientists, who, having generated new ideas, inspired the scientific community to create a significant number of claimed (highly cited) works. This being the case, the generators of ideas themselves have remained in the shade. Such scientists are often sought by leading research companies, but it is fairly difficult to discover them automatically owing to their low scientometric indices. At present, the identification of the latent initiators of the scientific mainstream can be formalized by revealing the scientists who simultaneously have a large value of the suggested citation index and a low value of the traditional citation index. The suggested index can only be used in practice based on automated accounting of scientific publications, for example, that based on the Web of Knowledge, Scopus, Google Scholar, or eLibrary.ru systems.

REFERENCES

1. Egghe, L., Mathematical theory of the h- and g-index in case of fractional counting of authorship, *J. Am. Soc. Inform. Sci. Technol.*, 2008, vol. 59, pp. 1608–1616.
2. Schreiber, M., A modification of the h-index: The h(m)-index accounts for multi-authored manuscripts, *J. Informetrics*, 2008, vol. 2, pp. 211–216.
3. Buela-Casal, G., Assessing the quality of articles and scientific journals: Proposal for weighted impact factor, *Psychology in Spain*, 2004, vol. 8, no. 1, pp. 60–76.
4. van Noorden, R., A profusion of measures, *Nature*, 2010, vol. 465, pp. 864–866.
5. Garcia-Perez, M.A., The Hirsch h index in a non-mainstream area: Methodology of the behavioral sciences in Spain, *The Spanish J. Psychol.*, 2009, vol. 12, pp. 833–849.
6. Soifer, V.N., International soros program for education. Part 2. Results of name concourses, *Soros. Obraz. Zh.,* 1996, no. 1, pp. 4–16.
7. Franceschini, F., Maisano, D., Perotti, A., and Proto, A., Analysis of the ch-index: An indicator to evaluate the diffusion of scientific research output by citers, *Scientometrics*, 2010, vol. 85, pp. 203–217.
8. Kara-Murza, S.G., Testing in science and approaches to scientific contribution estimation, *Vestnik Akad. Nauk SSSR,* 1981, no. 5, pp. 68–75.
9. Chen, C. and Hicks, D., Tracing knowledge diffusion, *Scientometrics*, 2004, vol. 59, pp. 199–211.
10. Folin, O. and Ciocalteu, V., On tyrosine and tryptophane determinations in proteins, *J. Biolog. Chem.*, 1927, vol. 73, pp. 627–650.
11. Lowry, O.H., Rosbrough, N.J., Farr, A.L., and Randall, R.J., Protein measurement with the Folin phenol reagent, *J. Biolog. Chem.*, 1951, vol. 193, pp. 265–275.
12. Gao, X. and Guan, J., Network model of knowledge diffusion, *Scientometrics*, 2012, vol. 90, pp. 749–762.
13. Schubert, A., Using the h-index for assessing single publications, *Scientometrics*, 2009, vol. 78, pp. 559–565.
14. Atallah, G. and Rodriguez, G., Indirect patent citations, *Scientometrics*, 2006, vol. 67, pp. 437–465.

Translated by L. Solovyova