ISSN 0147-6882, Scientific and Technical Information Processing, 2009, Vol. 36, No. 4, pp. 198–202. © Allerton Press, Inc., 2009.
Original Russian Text © V. M. Moskovkin, 2009, published in Nauchno-Technicheskaya Informatsiya, Seriya 1, 2009, No. 7, pp. 12–16.

The Potential of Using the Google Scholar Search Engine for Estimating the Publication Activities of Universities

V. M. Moskovkin

Received March 20, 2009

Abstract—This paper studies the potential of using the Google Scholar search engine for estimating the publication activities of universities and considers a procedure for such estimation with the help of queries for the English names of universities. The publication structures for 2008 have been built for ten selected universities of the world, including MSU. The publication activities of the universities under consideration in 2007, has been compared based on the citation database of the US Institute for Scientific Information (Web of Knowledge) and Google Scholar search engine (GS-publications).

Key words: *Google Scholar search engine, university publication activity, open access, citation indexes, publication structures.*

DOI: 10.3103/S0147688209040029

A cluster of publications connected with the use of the Google Scholar search engine for carrying out such investigations has appeared now in the foreign scientific literature devoted to the scientometric methods of investigations. It is noted that the databases of the US Institute for Scientific Information (ISI Citation Indexes) were a unique overall source of citation data until recently.

These databases have long been widely used abroad in scientific management, coping with some of their imperfections as well as the absence of other imperfections.

Two alternatives to these databases, i.e., the commercial Scopus search engine developed by the very large Elsevier Publishing House for scientific periodicals and open-access Google Scholar search engine have appeared comparatively recently [1].

Works [2, 3] show that Google Scholar covers a much greater quantity of documents compared to the databases of the US Institute for Scientific Information, thus making a great contribution to the movement towards open access to the results of scientific investigations.

Work [4] notes that the Google Scholar search engine provides a new method for discovering potentially relevant articles on these themes at the expense of identifying the articles cited in other works. Therefore, an important property of this search engine is that researchers can use it to track the mutual relationships between authors citing articles on a similar subject and also to determine the frequency with which other authors cite a specific article (with the "cited by" option). It also makes the conclusion that the Google Scholar search engine provides a free alternative and supplements other citation indexes.

The databases of the US Institute for Scientific Information index approximately one third of the total quantity of reviewed scientific journals, which number about 25 000 at present. This being the case, Google Scholar and Google Books index many more scientific documents, but still are not able to achieve complete coverage, since only 15% of the current annual scientific output is represented by OA publications [5].

Our review of scientometric investigations has shown that there are no works studying the publication structure and webometric estimates of university scientific outputs with the help of the Google Scholar search engine.

When studying this output, we paid attention to the impossibility of qualitatively obtaining it by measuring the responses to the URL addresses of university sites. The large quantity of irrelevant responses in the form of different administrative information (decisions of a scientific council, university administration, etc.) often appear for post-soviet universities. The situation arises for western universities when, for example, in the case of approximately equal publication activities of scientists from American universities (the Harvard and Chicago Universities) and British universities (Cambridge and Oxford Universities), the American scientists had more responses to queries for their URL-addresses[1], by an order of magnitude, although queries for the names

[1] Harvard.edu, site: uchicago.edu, site: ox.ac.uk, site: cam.ac.uk give 1 310 000, 60 400, 8090, and 9330 documents, respectively; the measurements were carried out by us at the beginning of January 2009.

of these universities gave advantages to Britain universities[2].

In our opinion, this is caused by the quality of the organization of information on a site. Thus, the multiplicity of responses to the Harvard University site is connected with the existence of a scientific internet magazine (sciencemag.org), and when a query is made to the Chicago University site, the first thousand responses that are shown by the Google Scholar search engine are for the articles of the excellently composed journal collection of the university ("Chicago Journals") that is found on the uchicago.press platform.

It is obvious that the sites of universities present by no means all publications of their scientists, and, as for post-soviet universities, the practice of placing scientific articles on their sites is completely nonexistent.

In connection with the above-stated facts, we decided to test the generally recognized English names of universities with the help of the Google Engine search engine, instead of testing their URL-sites with its use, as one Spanish cybermetric laboratory does when calculating the webometric ranking of the world's universities (www.webometrics.info). The experiments with the leading universities of the world showed the high relevance of this search. In the first place, Google Scholar finds articles placed on the online platforms of the largest publishing houses, such as Elsevier, Springer, Blackwell, Wiley, etc., i.e., "convertible" articles included in the databases of the US Institute for Scientific Information. In addition, this search engine efficiently finds articles from online journals and open-access university repositories.

Let us also note that Google Scholar also includes a small percentage of scientific monographs provided by Google Books in the results of its search.

We have already been able to show that the relevance of an advanced search with an exact phrase increases in the following order: in the absence of restrictions on the fields of science and time intervals –> with the assignment of fields of sciences –> with the simultaneous assignment of fields of science and time intervals of search.

Besides the total quantity of articles in a given knowledge field (7 fields) that are obtained in response to a query for the English name of a specific university, Google Scholar gives the values of the total number of citations to each article with the opportunity of browsing through the names of scientific works citing this article (with the help of the "by cited" option). Our contacts with the Google Scholar team showed that a procedure permitting one to summarize the citations in the entire assembly of found articles was still nonexistent, but the Google Scholar team received the idea of developing this procedure with interest. If it is realized, this will provide the opportunity to calculate the full-value webometric ranking of the scientific-publication activities of the world's universities. When this ranking is calculated, the problem of identifying all the generally accepted names of universities arises. For example, generally accepted French and English names must be used for the universities of the French-speaking Canadian provinces; all the main foreign-language names of the European universities of the non-English-speaking countries must be used for them, as well as their English names. The permanent process of renaming universities should be taken into consideration for post-soviet countries. When working with the Google Scholar search engine, we noted the fluctuation of responses to queries for the names of universities, which is connected with the possible temporary absence of access, exclusion of duplications and irrelevant responses. Therefore, in our opinion, it is expedient to use smoothing procedures (to calculate the average temporal trend) when calculating a resultant webometric index.

We believe that with time, as the creation of open-access university repositories becomes more intensive, the probability of duplicated responses to queries for the URL-addresses of university sites will increase, since previously unpublished articles will be placed (self-archived) in such repositories (mainly, in the form of an author's PDF-files).

It is difficult to say in advance how effectively the Google Scholar search engine will cope with the increasing scale of article duplication.

Nine foreign universities that occupied the highest positions according to the quantity of published articles (included in the SSI and SSCI databases of the US Institute for Scientific Information) in the 2008 Chinese and Taiwanese rankings of universities were chosen by us as experiments with the Google Scholar search engine. The leading Russian post-soviet university, the MSU, was chosen for comparison. Tables 1 and 2 show the publication structures for 2008 for these universities in the quantitative and percentage terms, as well as the enlarged publication structure. The caps of these tables give the main names of the universities, according to which an advanced search with exact phrase was performed. The inverted names of these universities (for example, the University of Chicago—the Chicago University) were also taken into account in the queries of the Google Scholar, with the exception of the Hopkins University, University of California, University of Tokyo, and Moscow State University. The greatest shares of responses for the inverted names were observed for the Chicago, Cambridge, and Oxford universities. The responses to the inverted name of the University of Tokyo often brought to other universities (the Tokyo University of Agriculture, Technology, or Science); therefore, they were not taken into account in summary estimates. The data of table 2 were calculated based on the percentage distribution of the data given in table 1. For example, the share of publications in the field of life sciences for Harvard University amounted

[2] Advanced search with an exact phrase.

Table 1. The publication structure for the selected largest universities in the world in 2008, obtained using the Google Scholar search engine on January 22, 2009

Fields of sciences	Stanford University	Harvard University	Columbia University	University of California-Berkeley	Johns Hopkins University	Chicago University	Cambridge University	Oxford University	University of Tokyo	Moskow State University
1. Biology, Life Science, and Environmental Science	2346/8.8	3102/9.7	2977/12.2	1670/14.2	2240/10.1	4453/8.6	17920/13.1	15100/12.3	2460/19.3	560/14.8
2. Business, Administration, Finance, and Economics	2686/10.1	5802/18.1	2210/9.0	1090/9.3	1470/6.7	6850/13.3	11759/8.6	10520/8.6	376/2.9	57/1.5
3. Chemistry and Materials Science	1118/4.2	822/2.5	935/3.8	980/8.3	584/2.6	736/1.4	7340/5.3	8167/6.7	1720/13.5	957/25.2
4. Enginering, Computer Science, and Mathematics	4455/16.7	2161/6.7	1680/6.9	2370/20.1	2200/9.9	1914/3.7	20230/14.7	9703/7.9	1910/15.0	583/15.4
5. Medicine, Pharmacology, and Veterinary Science	5236/19.7	1832/5.7	4391/18.0	698/5.9	5180/23.5	4870/9.4	9670/7.1	19970/16.3	2320/18.2	71/1.9
6. Physics, Astronomy, and Planetary Science	2112/7.9	1591/5.0	1523/6.2	1950/16.6	1630/7.4	4017/7.8	19190/14.0	7290/6.0	3260/25.6	1380/36.4
7. Social Sciences, Arts, and Humanities	8682/32.6	16813/52.3	10729/43.9	3010/25.6	8780/39.8	28830/55.8	50980/37.2	51570/42.2	703/5.5	180/4.8
Total	26635/100	32123/100	24445/100	11768/100	22084/100	51670/100	137089/100	122320/100	12749/100	3788/100

Note: The numerator gives the number of publications, and the denominator gives %.

Table 2. The enlarged publication structure for the selected largest universities in the world in 2008, obtained using the Google Scholar search engine on January 22, 2009

Enlarged fields of sciences	Stanford University	Harvard University	Columbia University	University of California-Berkeley	Johns Hopkins University	Chicago University	Cambridge University	Oxford University	University of Tokyo	Moskow State University
Natural and techical sciences	28.8	14.2	16.9	45.0	19.9	12.9	34.0	20.6	54.1	77.0
Life sciences	28.5	15.4	30.2	20.1	33.6	18.0	20.2	28.6	37.5	16.7
Socio-economic sciences and humanities. Art.	42.7	70.4	52.9	34.9	46.5	69.1	45.8	50.8	8.4	6.3
Total	100	100	100	100	100	100	100	100	100	100

to 9.7 + 5.7 = 15.4%. Table 2 shows that the scientific schools of the socioeconomic and humanitarian fields are predominant at Harvard and the Chicago University. The opposite picture is observed for the Universities of California, Tokyo, and Moscow. Scientific schools in the field of life sciences are most heavily represented at the University of Tokyo and Johns Hopkins University. The post-soviet publication structure represented by the scientific output of the MSU is characterized by the clear predominance of "convertible" natural-scientific and technical publications and, consequently, scientific schools of the natural scientific and technical directions (with the exception of medical and biological scientific schools).

Let us compare now the publication activities of the universities under consideration that was obtained based on the citation databases of the US Institute for Scientific Information (ISI) and Google Scholar search engine (GS). For this purpose, we addressed ourselves to the Taiwanese Ranking of Scientific Papers for World Universities. This ranking contains the "Current Articles" index, which represents the annual quantity of publications obtained based on the SCI and SSCI databases (Thomson–Reuter). This index in the 2008 ranking of world universities was calculated for 2007. Its maximal value, taken as 100% was for Harvard University and was equal to 11 221 articles [3]. Its absolute values were recalculated by us for the remaining universities based on the maximal value of this index. The quantities of scientific articles that were obtained by the previously described method with the help of the Google Scholar search engine were also calculated for the same year (GS-publications in table 3). The surplus of the webometric index of the university publication activities over its traditional index was calculated as well in table 3. As is clear, this ratio varies rather strongly. Meanwhile, it is logical to suppose that the ratio of the total number of publications to "convertible" publications (Thomson–Reuter) is an approximately constant value for different universities, i.e., there must be a good linear correlation between these indices. The absence of such a correlation between the indices of Thomson–Reuter and GS-publications speaks only for the bad Web-presentation of publications for universities for which the ratio "GS-publications/Thomson–Reuter" is low.

The index of Thomson–Reuter must be included in articles from the A&HCI database for more correct calculations, as the Google Scholar search engine covers such articles.

It is to be noted that the PUB index completely corresponds to the Current articles index (the Taiwanese Ranking) in the Shanghai Ranking of World Universities, but it cannot be directly used to calculate the absolute values of university publications included in the SCI and SSCI databases, since a coefficient of 2 was used for socioeconomic articles.

Consequently, we have shown the possibility of quantitatively estimating the publication activities of universities with the help of the Google Scholar search engine, confirmed the results of foreign researchers on the wider coverage of scientific publications by this search engine in comparison with the databases of the US Institute for Scientific information, and built the publication structures for ten selected examples of the leading world universities. The further development of this approach must follow the path of separating book publications (the mark "Book") and citations (the mark "Citation") in the responses of the Google Scholar search engine, despite the large percentage of these responses. However, this work, together with the calcu-

[3] The absolute value of the "Current articles" index was kindly furnished to us by Ru-rong Hsiao (the Chief of the Performance Evaluation Section HEEACT of Taiwan.

Table 3. The publication activity of the largest universities in the world obtained based on the data of the Taiwanese Ranking of World Universities and Google Scholar search engine, 2007

Universities	The quantity of articles			GS-publications/ Thomson-Reuter
	Thomson-Reuter		GS-publications	
	%	the absolute value		
Harvard University	100	11221	46768	4.2
University of Tokyo	62.51	7014	15495	2.2
Johns Hopkins University	52.98	5878	27124	4.6
University of California-Berkeley	47.67	5349	14571	2.7
Stanford University	47.87	5372	35320	6.6
Columbia University	43.10	4836	32990	6.8
Oxford Universit	39.60	4444	142344	32.0
Cambridge University	39.35	4416	173060	39.2
Chicago University	34.78	3903	73016	18.7
Moskow State University	28.05	3148	5021	1.6

lation of the total number of citations in all the found academic documents (the "by cited" option) can be done only in cooperation with the Google Scholar team.

REFERENCES

1. Bar-Ilan, J., Which h-index? – A Comparison of WoS, Scopus and Google Scholar, */Scientometrics*, 2008, Vol. 74, no 2, pp. 257–271.
2. Kousha, K., Thelwall, M., Google Scholar Citations and Google Web/URL Citations: A Multi-Discipline Exploratory Analysis, *Journal of the American Society for Information Science and Technology*, 2007,Vol. 58, no 7, pp. 1055–1065.
3. Kousha, K., Thelwall, M., Sources of Google Scholar Citation outside the Science Citation Index: A comparison between Four Science Disciplines, *Scientometrics*, 2008, Vol. 74, no 2, pp. 273–294.
4. Noruzi, A., Google Scholar: The New Generation of Citation Indexes, *Libri*, 2005, Vol.55, no 4, pp. 170–180.
5. Brody, T., Carr, L., Gingras, Y., Hajjem, Ch., Harnad, S., Alma Swan. Incentivizing the Open Access Research Web Publication – Archiving, Data-Archiving and Scientometrics, *CTWatch Quarterly*, 2007 (August).