The calculation of the single publication *h* index and related performance measures

A web application based on Google Scholar data

Andreas Thor
Department of Computer Science, University of Leipzig, Leipzig, Germany, and
Lutz Bornmann
Eidgenössische Technische Hochschule, Zurich, Switzerland

Refereed article received 18 April 2010
Approved for publication 5 August 2010

Abstract

Purpose – The single publication h index has been introduced by Schubert as the h index calculated from the list of citing publications of one single publication. This paper aims to look at the calculation of the single publication h index and related performance measures.

Design/methodology/approach – In this paper a web application is presented where the single publication h index and related performance measures (the single publication m index, h^2 lower, h^2 centre, and h^2 upper) can be automatically calculated for any publication indexed by Google Scholar.

Findings – The use of the application is demonstrated by means of the citation performance of two publications.

Originality/value – To the authors' knowledge this web application is the first instrument to automatically calculate the single publication h index and related performance measures based on Google Scholar data. This is a new service especially from the perspective of the related performance measures.

Keywords Publications, Indexing, Computer applications, Search engines

Paper type Research paper

Introduction

Hirsch (2005) has proposed the h index as a criterion to quantify the scientific output of a single researcher. "The automatic calculation of h-indices has even become a built-in feature of major bibliographic databases such as Web of Science and Scopus" (van Eck and Waltman, 2008, p. 263). In recent years it has been proposed to use the h index not only for the performance measurement of single scientists, but also of journals, research groups, departments, and countries (see Bornmann and Daniel, 2007, 2009; Egghe, 2010). Schubert (2009) suggests calculating the h index for the citing publications of one single publication. This results in the "single publication h index". To calculate this index the citing publications of one single publication are gathered in a publication list. Next the citation counts are added to each (citing) publication in that list. The single publication h index for the publication in question is that number of (citing) publications in the list with citation counts $\geq h$.

Online Information Review
Vol. 35 No. 2, 2011
pp. 291-300

1468-4527
DOI 10.1108/14684521111128050

Schubert (2009, p. 560) justifies his proposal to calculate an *h* index for single publications as follows:

> Citation indicators usually measure the "direct impact" of publications, i.e., the amount of the citations received (whether in the form of simple counts, weighted sums or normalised units). Undoubtedly, however, publications may exert influence also indirectly, e.g., through their presence in reference lists [. . .] It seems therefore reasonable to construct indicators that take into account not only the direct [but] also the indirect citation influence of publications.

For a highly cited publication in particular the additional consideration of indirect citation influence should result in a more refined performance picture than with the use of bare citation counts. In this paper we present a web application to calculate the single publication *h* index with Google Scholar (GS) data (provided by Google, headquartered in Mountain View, California).

In recent years several disadvantages of the h index have been pointed out which are also valid for the single publication h index (e.g. additional citations of an already highly-cited publication mostly do not affect the h index value). This has led to the development of numerous variants of the *h* index (e.g. the *a* index, *m* quotient, *m* index, and *g* index) (for an overview see Bornmann and Daniel, 2009). The results of Bornmann *et al.* (2008, 2009a, b) show that regarding the *h* index and its variants, we are dealing with two types of indices. One type describes the most productive core within a publication list in terms of citation performance (e.g. the *h* index) and gives the number of publications in that core. The other type of index describes the impact of the publications in the core (e.g. the *m* index: the median number of citations received by publications in the Hirsch core – this is the publications ranking smaller than or equal to *h*). For evaluative purposes Bornmann *et al.* (2008) propose the use of a combination of two indices, where one index relates to the first index type and the other index relates to the second type. Against this backdrop our web application not only calculates the single publication *h* index, but also the single publication *m* index.

Bornmann *et al.* (2010) introduce an approach providing additional information to the *h* index: h^2 lower, h^2 centre, and h^2 upper. As the results of Bornmann *et al.* (2010) show, scientists with similar *h* index values may have very different research performances. Their approach allows the quantification of three areas within a citation distribution: the low impact area (h^2 lower), the area captured by the *h* index (h^2 centre), and the area of publications with the highest visibility (h^2 upper). The *h* index refers to the area h^*h and normally captures only a small part of the publication and citation data in a publication list if the distribution is right-skewed. The *h* index does not take into consideration the areas starting at *h* citations (h^2 upper) or starting at *h* publications (h^2 lower). For this reason the area proportions h^2 lower, h^2 centre, and h^2 upper are provided as additional information to the single publication *h* index and *m* index by our web application. A high percentage for h^2 upper indicates a publication list dominated by highly cited publications. A high percentage for h^2 lower indicates a relatively large number of publications of little impact in the list.

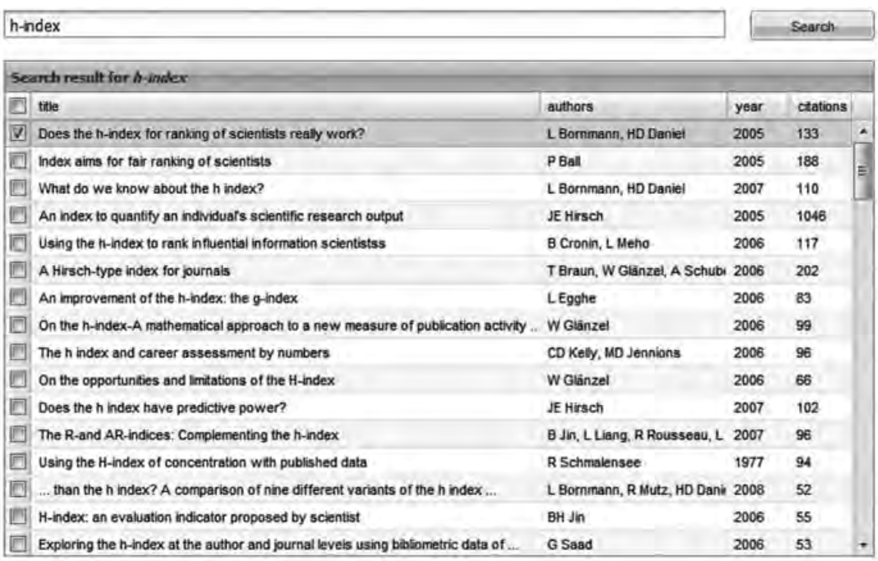
Description of the web application

The web application can be accessed at http://labs.dbs.uni-leipzig.de/gsh. It is built on top of GS, i.e. it automatically retrieves the relevant GS data during run-time. This application is free and available to everyone.

To find a publication in GS for the calculation of the single publication h index and the related performance measures, relevant keywords from the title of a publication, the author and/or journal names are entered into the search field. As an example we use a publication of one of the authors entitled "Does the h-index for ranking scientists really work?" which appeared in *Scientometrics* in 2005. By using the key word "h index" a search in GS results in a list of more or less matching publications (see Figure 1). For each found publication the title, the authors, the publication year, and the citation counts are displayed.

To get the single publication h index and the related measures, the publication in question is marked and then the button "Get h index" is pressed. As a result the single publication h index, h^2 lower, h^2 centre, h^2 upper, the single publication m index, and the total citation counts (#citations) are displayed (see Figure 2). The bottom part of the webpage comprises a citation distribution graph of the citing publications. To this end the citing publications are rank-ordered by their citation numbers and the number of citations is plotted for each citing publication. In addition the graph visualises the area proportions h^2 lower, h^2 centre, and h^2 upper. With a percentage of 53.3 percent for h^2 upper it is clearly shown in Figure 2 that the citing publications of the publication in question are dominated by publications with high citation counts.

Since citations for a single publication can consist of external- and self-citations, our web application offers the option to exclude self-citations from further computations. The user may therefore uncheck the "incl. self-citations" box and rerun the analysis (see Figure 3). All citing publications of the publication in question sharing at least one author with that publication are filtered away for the calculation of the performance indicators. Figure 3 shows that the number of citations has decreased from 133 to 121. The performance measures are calculated on the basis of these 121 citing publications and their citations (self- and external-citations). That means only self-citations of the



Note: Date of search: 15 April 2010

Figure 1. Search results for "h-index" based on a Google Scholar search

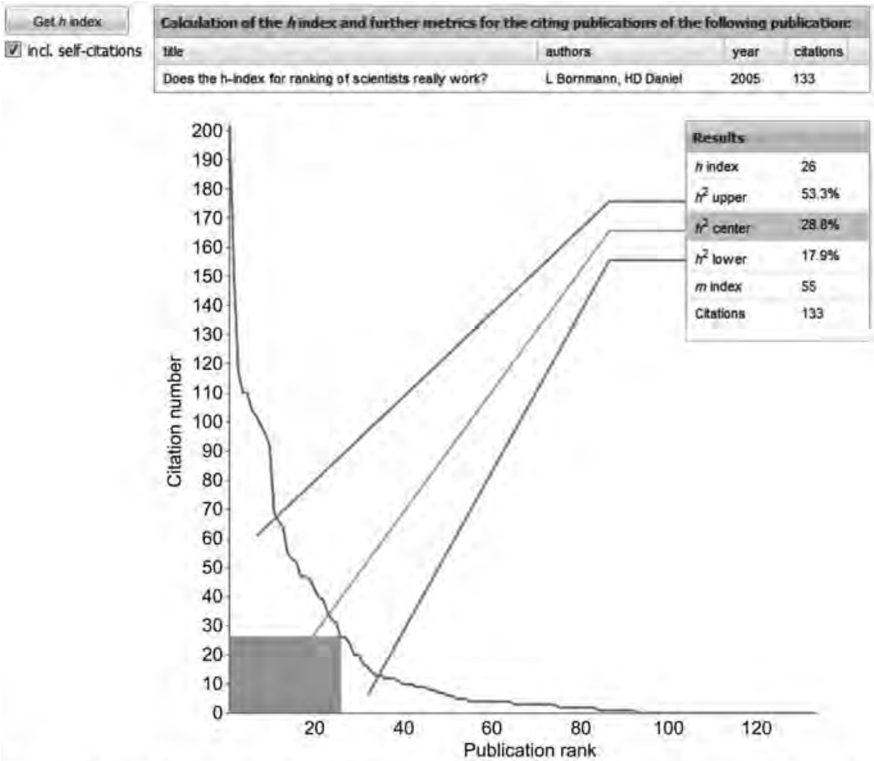


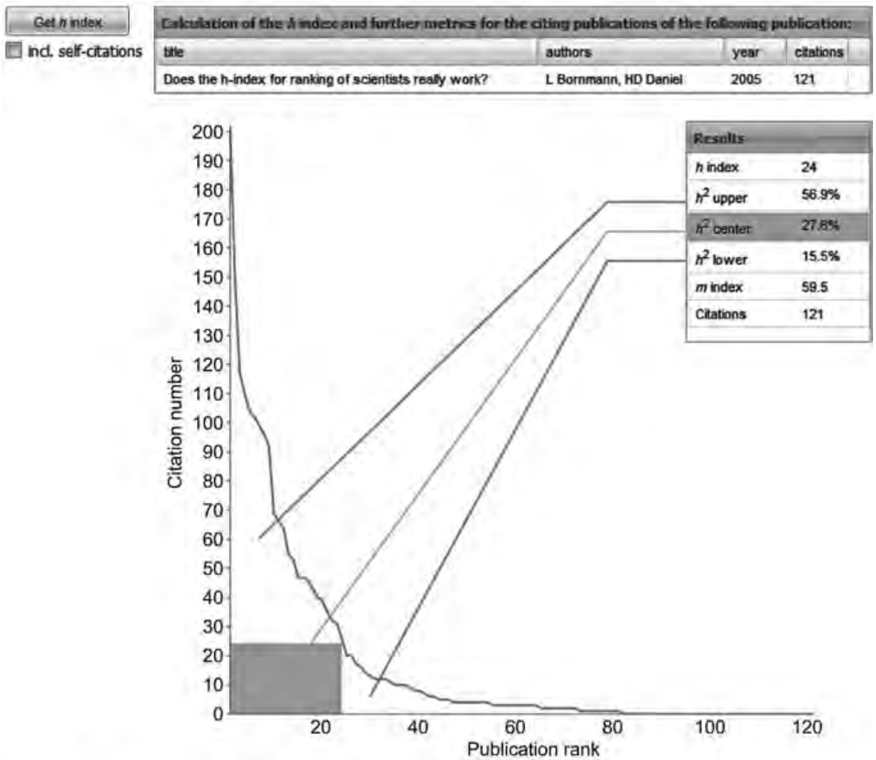
Figure 2.
Analysis result for the publication "Does the *h*-index for ranking of scientists really work?" by Bornmann and Daniel (2005), including self-citations

Notes: The result comprises *h* index, *h*2 upper, *h*2 centre, *h*2 lower, *m* index, and the total number of citations. Additionally the citation distribution of the rank-ordered citing publications is plotted. Date of analysis: 15 April 2010

Source: Bornmann and Daniel (2005)

publication in question are eliminated whereas all citations (self- and external-citations) of the citing publications are used for the computation of the *h* index and the related measures.

The web application also addresses the GS data quality problems. In particular GS often contains duplicates. For example Figure 4 shows a search result for "allintitle:merge purge large" consisting of eight GS records that all refer to the same publication (that ironically deals with the automatic identification of duplicates) (Hernández and Stolfo, 1995). The user of our web application may therefore mark all duplicates of a publication and then perform a joint citation analysis. To this end the web application retrieves the lists of citing publications for each selected GS record individually and subsequently merges them into a combined list. In doing so the web application ensures that all relevant citations are taken into account even if they refer to different GS records. The removal of duplicates can significantly influence the computation of the single publication *h* index and related measures. Figure 5 shows the result for the top record of the search displayed in Figure 4; Figure 6 shows the combined results for all duplicates of Hernández and Stolfo (1995). The use of all



Notes: The elimination of self-citations leads to slightly different results (see Figure 2 for comparison). Date of analysis: 15 April 2010

Source: Bornmann and Daniel (2005)

Figure 3.
Analysis result for the publication "Does the *h*-index for ranking of scientists really work?" by Bornmann and Daniel (2005), excluding self-citations

duplicates leads to a higher number of citations (528 vs 492) and to a slightly higher *h* index value (55 vs 54).

Discussion

GS has been available on the internet since 2004 as a beta release, and it is particularly interesting for conducting citation analyses, because in contrast to the other databases it can be accessed for free (Neuhaus and Daniel, 2008). According to Harzing and van der Wal (2008, p. 62):

> [. . .] an important practical advantage of GS is that it is freely available to anyone with an Internet connection and is generally praised for its speed. On the other hand, the WoS (Web of Science, provided by Thomson Reuters, Philadelphia) is only available to those academics whose institutions are able and willing to bear the (quite substantial) subscription costs of the WoS and other databases in Thomson ISI's Web of Knowledge.

In addition GS does not search only peer-reviewed and professional research journals (as WoS does, for example):

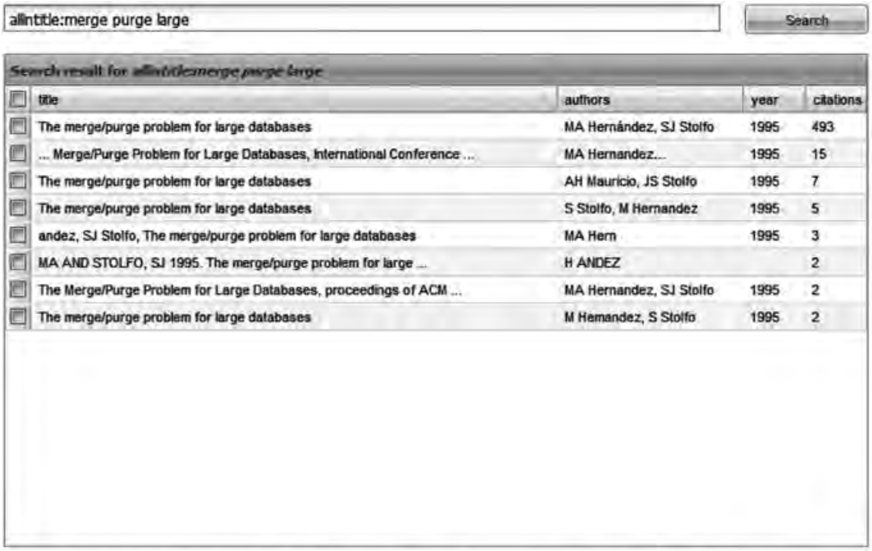


Figure 4.
Search result for "allintitle:merge purge large"

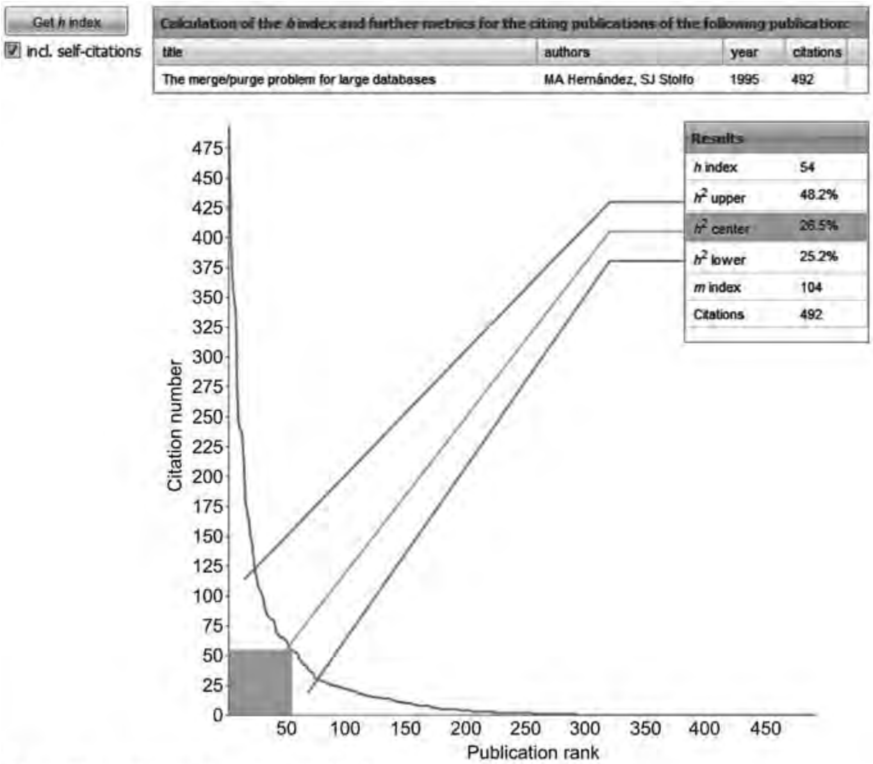
Notes: All retrieved Google Scholar records refer to the same publication by Hernández and Stolfo. Date of search: 15 April 2010

Source: Hernández and Stolfo (1995)

> It searches lots of non-traditional sources, including preprint archives, conference proceedings and institutional repositories, often locating free versions of articles on author websites (Giles, 2005, p. 554).

Because of the fee-based databases' poor coverage of certain fields and the difficulty of citation analysis of publications that are not published in journals indexed by Thomson Reuters (or in Scopus, provided by Elsevier, Amsterdam), the analysis of GS data can be a great advantage. The calculation of the single publication h index and the related performance measures based on GS data by using our web application offers interesting insights into the citation performance not only of journal papers, but also of any other publication type (e.g. books, book chapters).

However, independently of the field or discipline, anyone using GS must be aware that the database is still in beta testing (Bar-Ilan, 2008). According to an overview by Bar-Ilan (2008) neither the Boolean operators, nor the range operator (for limiting the date of publication) work properly. Furthermore it is not always possible "to correctly identify the publication year of the item, and citations are not always attributed to the correct publication" (Bar-Ilan, 2008, p. 260). For Jacsò (2008a) GS "does a really horrible job matching cited and citing references" and "often can't tell apart a page number from a publication year, part of the title of a book from a journal name, and dumps at you absurd data" (see also Perkel, 2005). Jacsò (2010) reviews the recent developments in Google Scholar's management of bibliographic metadata and acknowledges the usefulness of GS's keyword search but also illustrates GS data quality problems such as phantom authors and phantom publication years. In addition "the hit counts and the citation counts of Google Scholar keep changing



Note: Date of analysis: 15 April 2010

Figure 5.
Analysis result for the top record of the search displayed in Figure 4

dramatically. If they were increasing, it could be chalked up to adding new records, but often these counts decrease because of deleting records from the database" (Jacsò, 2008b, p. 270). Given these weaknesses it seems justified that Gardner and Eng (2005) conclude that Google should improve GS significantly in the beta testing phase before it becomes fully operational.

The reason for the substandard GS data quality lies in the automatic generation of the GS data set (among other things, automatic extraction of reference lists from PDF files), which may lead to both heterogeneous bibliographic information for the same publication (e.g. due to missing authors, authors listed in incorrect order, differences in the names used for journals and conferences) and errors in the metadata (e.g. due to typographical errors in titles, extraction errors when splitting reference strings). The performance measures calculated by our web application may not be influenced as significantly by the questionable data quality as other measures, such as the journal impact factor (JIF (provided by Thomson Reuters)). The bibliographic information of the citing publications are of little importance for the computation of the single publication h index (and the related measures) because only the citation numbers of the citing publications are considered. This is not the case for other metrics such as the JIF

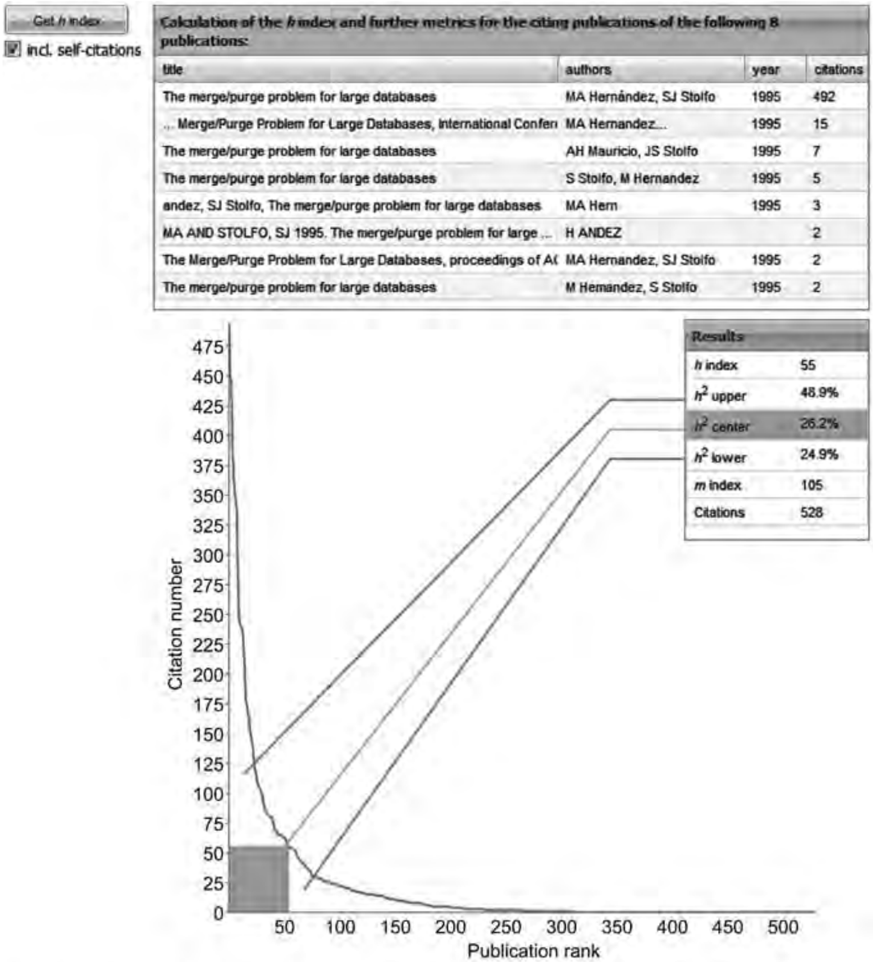


Figure 6.
Joint analysis result for all publication records shown in Figure 4

Notes: The consideration of all duplicates of the same real-world publication influences the results of the analysis, e.g. the *h* index. Date of analysis: 15 April 2010

which relies on the publication year of the citing publications. However missing citations and phantom citations in GS may of course influence the computation of the *h* index (and related measures) even though the *h* index is quite robust.

GS's automatic generation of bibliographic information leads to duplicates for the same publication. The identification and removal of all duplicates is obviously crucial with regard to the calculation of the single publication *h* index and the related performance measures. Automatic identification of duplicate records is a very challenging task receiving a lot of attention in computer science research (Thor and Rahm, 2007). At present 100 percent correct automatic identification of duplicates is not possible. Therefore the users of our web application have to manually select all duplicates before calculating the performance measures.

All in all, the citation performance measuring for one single publication by using our web application based on GS data has advantages and disadvantages for the user. Because of the disadvantages we recommend that the user of the application checks the analysis results carefully and tests – whenever possible – the convergent validity (Bornmann *et al.*, 2009a, b) of the results by comparing it with the findings produced by other databases (e.g. WoS).

References

Bar-Ilan, J. (2008), "Which *h*-index? A comparison of WoS, Scopus and Google Scholar", *Scientometrics*, Vol. 74 No. 2, pp. 257-71.

Bornmann, L. and Daniel, H.-D. (2005), "Does the *h* index for scientists really work?", *Scientometrics*, Vol. 65 No. 3, pp. 391-2.

Bornmann, L. and Daniel, H.-D. (2007), "What do we know about the *h* index?", *Journal of the American Society for Information Science and Technology*, Vol. 58 No. 9, pp. 1381-5.

Bornmann, L. and Daniel, H.-D. (2009), "The state of *h* index research. Is the *h* index the ideal way to measure research performance?", *EMBO Reports*, Vol. 10 No. 1, pp. 2-6.

Bornmann, L., Mutz, R. and Daniel, H.-D. (2008), "Are there better indices for evaluation purposes than the *h* index? A comparison of nine different variants of the *h* index using data from biomedicine", *Journal of the American Society for Information Science and Technology*, Vol. 59 No. 5, pp. 830-7.

Bornmann, L., Mutz, R. and Daniel, H.-D. (2010), "The *h* index research output measurement: two approaches to enhance its accuracy", *Journal of Informetrics*, Vol. 4 No. 3, pp. 407-14.

Bornmann, L., Mutz, R., Daniel, H.-D., Wallon, G. and Ledin, A. (2009a), "Are there really two types of *h* index variants? A validation study by using molecular life sciences data", *Research Evaluation*, Vol. 18 No. 3, pp. 185-90.

Bornmann, L., Marx, W., Schier, H., Rahm, E., Thor, A. and Daniel, H.-D. (2009b), "Convergent validity of bibliometric Google Scholar data in the field of chemistry. Citation counts for papers that were accepted by Angewandte Chemie International Edition or rejected but published elsewhere, using Google Scholar, Science Citation Index, Scopus, and Chemical Abstracts", *Journal of Informetrics*, Vol. 3 No. 1, pp. 27-35.

Egghe, L. (2010), "The Hirsch index and related impact measures", *Annual Review of Information Science and Technology*, Vol. 44, pp. 65-114.

Gardner, S. and Eng, S. (2005), "Gaga over Google? Scholar in the social sciences", *Library Hi Tech News*, Vol. 2 No. 8, pp. 42-5.

Giles, J. (2005), "Science in the web age: start your engines", *Nature*, Vol. 438 No. 7068, pp. 554-5.

Harzing, A.W. and van der Wal, R. (2008), "Google Scholar as a new source for citation analysis", *Ethics in Science and Environmental Politics*, Vol. 8 No. 1, pp. 61-73.

Hernández, M.A. and Stolfo, S.J. (1995), "The merge/purge problem for large databases", in Carey, M.J. and Schneider, D.A. (Eds), *Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data*, ACM Press, New York, NY, pp. 127-38.

Hirsch, J.E. (2005), "An index to quantify an individual's scientific research output", *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 102 No. 46, pp. 16569-72.

Jacsò, P. (2008a), "Google Scholar and the scientist", available at: www2.hawaii.edu/~jacso/extra/gs/ (accessed 5 June 2008).

Jacsò, P. (2008b), "The plausibility of computing the *h*-index of scholarly productivity and impact using reference-enhanced databases", *Online Information Review*, Vol. 32 No. 2, pp. 266-83.

Jacsò, P. (2010), "Metadata mega mess in Google Scholar", *Online Information Review*, Vol. 34 No. 1, pp. 175-91.

Neuhaus, C. and Daniel, H.-D. (2008), "Data sources for performing citation analysis – an overview", *Journal of Documentation*, Vol. 64 No. 2, pp. 193-210.

Perkel, J.M. (2005), "The future of citation analysis", *The Scientist*, Vol. 19 No. 20, p. 24.

Schubert, A. (2009), "Using the *h*-index for assessing single publications", *Scientometrics*, Vol. 78 No. 3, pp. 559-65.

Thor, A. and Rahm, E. (2007), "MOMA – a mapping-based object-matching system", *3rd Biennial Conference on Innovative Data Systems Research, Asilomar, CA, 7-10 January*, available at: www.cidrdb.org (accessed 21 December 2010).

van Eck, N.J. and Waltman, L. (2008), "Generalizing the *h*- and *g*-indices", *Journal of Informetrics*, Vol. 2 No. 4, pp. 263-71.

About the authors

Andreas Thor received a diploma and a PhD in Computer Science in 2002 and 2008, respectively, from the University of Leipzig, Germany. He is a research scientist with the database group in Leipzig. He is currently a visiting research scientist at the University of Maryland Institute for Advanced Computer Studies. His research areas deal with the integration of web data sources. More specifically, he has been working on approaches for entity resolution, ontology alignment, and flexible integration architectures. Andreas Thor is the corresponding author and can be contacted at: thor@informatik.uni-leipzig.de

Lutz Bornmann (www.researcherid.com/rid/A-3926-2008) is a Researcher at the Professorship for Social Psychology and Research on Higher Education of the ETH (Swiss Federal Institute of Technology) Zurich. Since the late 1990s he has been working on issues around the promotion of young academics and scientists in the sciences and on quality assurance in higher education. His current academic interests include research evaluation, peer review and bibliometric indicators, especially the *h* index. He is a member of the Editorial Board of the *Journal of Informetrics* (Elsevier) and of the Advisory Editorial Board of *EMBO Reports* (Nature Publishing Group).