Manuscript accepted for publication in Online Information Review 35(2)

The calculation of the single publication *h* index and related performances measures: A Web application based on Google Scholar data

Andreas Thor & Lutz Bornmann

First author and corresponding author:
Dr. Andreas Thor
University of Leipzig
Department of Computer Science
PF 100920
D-04009 Leipzig
Tel.: +49 341 97 32 241
Fax: +49 341 97 32 309
E-mail: thor@informatik.uni-leipzig.de

Brief professional biography: Dr. Andreas Thor received a Diploma and a Ph.D. in Computer Science in 2002 and 2008, respectively, from the University of Leipzig, Germany. He holds an appointment as Research Scientist with the database group in Leipzig. Andreas is currently a visiting research scientist at University of Maryland Institute for Advanced Computer Studies (UMIACS). Andreas' research areas deal with integration of web data sources. More specifically, he has been working on approaches for entity resolution, ontology alignment, and flexible integration architectures.

Second author:
Dr. Lutz Bornmann
ETH Zurich
Professorship for Social Psychology and Research on Higher Education
Zähringerstr. 24
CH-8092 Zurich
Tel.: +41 44 632 48 25
Fax: +41 44 632 12 83
E-mail: bornmann@gess.ethz.ch

Brief professional biography: Dr. Lutz Bornmann (http://www.researcherid.com/rid/A-3926-2008) is a researcher at the Professorship for Social Psychology and Research on Higher Education of the ETH Zurich (Switzerland). Since the late 1990s, he has been working on issues in 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).

Abstract

Purpose: The single publication *h* index has been introduced by Schubert (2009) as the *h* index calculated for the list of citing publications of one single publication.

Approach: In this paper, we present a Web application where the single publication *h* index and related performance measures (the single publication *m* index, h^2 lower, h^2 center, 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: To our knowledge, our Web application is the first instrument to calculate automatically the single publication h index and related performance measures based on Google Scholar data. This service is a real novelty especially from the perspective of the related performance measures.

Key words

Single publication *h* index; Web application; Google Scholar

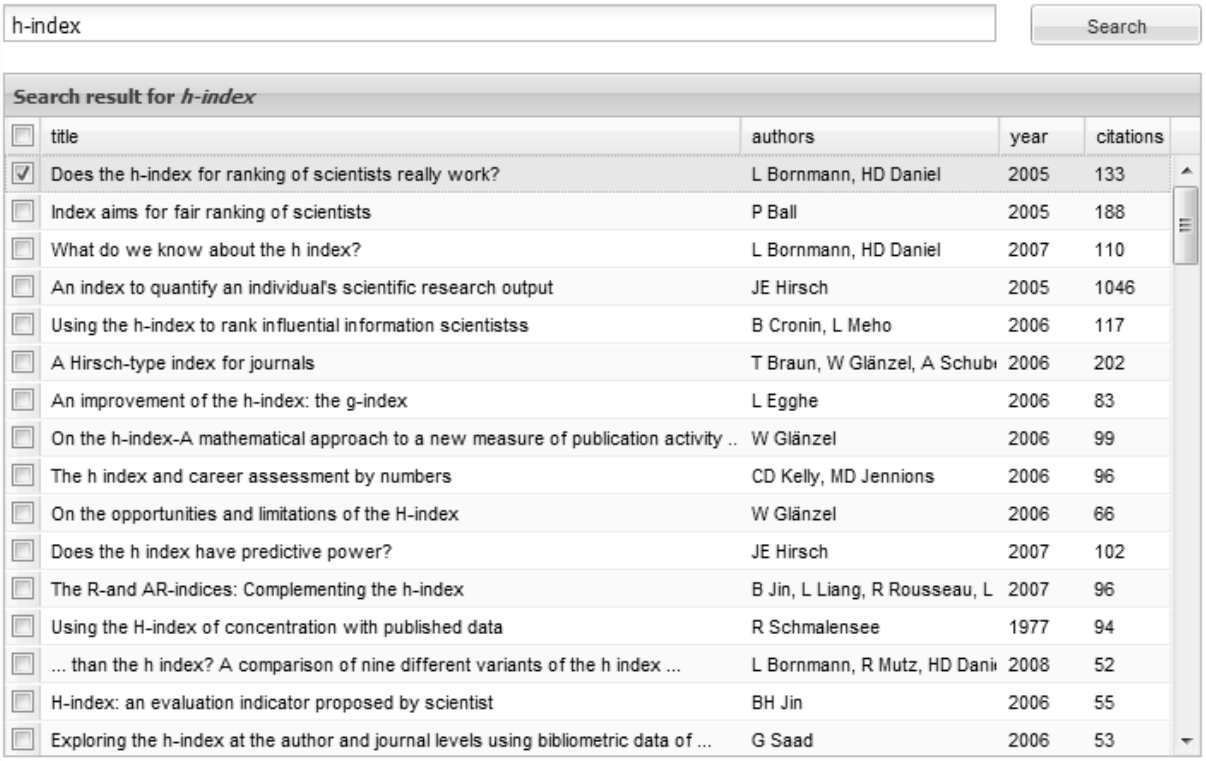
Introduction

Jorge 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 & Waltman, 2008, p. 263). In recent years, it has been proposed to use the *h* index not only for the performances measurement of single scientists, but also of journals, research groups, departments, and countries (see here Bornmann & Daniel, 2007; Bornmann & Daniel, 2009; Egghe, 2010). Schubert (2009) suggests to calculate the *h* index for the citing publications of one single publication. This results in the so-called “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 ≥ *h*.

Schubert (2009) 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 normalized 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” (p. 560). Especially for a highly cited publication 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, Inc., 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., it is insensible for highly-cited publications in a publication list). This has led to the development of numerous variants of the *h* index (e.g., *a* index, *m* quotient, *m* index, and *g* index) (see an overview in Bornmann & Daniel, 2009). The results of Bornmann, Mutz, and Daniel (2008) and Bornmann, Mutz, Daniel, Wallon, and Ledin (2009) 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 indices 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, Mutz, and Daniel (2008) propose the use of a combination of two indices, where one index relates to the one index type and the other index relates to the other type. Against this backdrop, our Web application does not only calculate the single publication *h* index, but also the single publication *m* index.

Bornmann, Mutz, and Daniel (2010) introduce an approach providing additional information to the *h* index: h^2 lower, h^2 center, and h^2 upper. As the results of Bornmann, Mutz, and Daniel (2010) show scientists with similar *h* index values may be very different research performance types. 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 center), and the area of publications with the highest visibility (h^2 upper). The *h* index refers to the area *h***h* and captures normally 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 center, and h^2 upper are provided as additional information to the single publication *h* index and *m* index by our Web application.


h-index

Search

Search result for *h-index*

	title	authors	year	citations
☑	Does the h-index for ranking of scientists really work?	L Bornmann, HD Daniel	2005	133
☐	Index aims for fair ranking of scientists	P Ball	2005	188
☐	What do we know about the h index?	L Bornmann, HD Daniel	2007	110
☐	An index to quantify an individual's scientific research output	JE Hirsch	2005	1046
☐	Using the h-index to rank influential information scientistss	B Cronin, L Meho	2006	117
☐	A Hirsch-type index for journals	T Braun, W Glänzel, A Schub	2006	202
☐	An improvement of the h-index: the g-index	L Egghe	2006	83
☐	On the h-index-A mathematical approach to a new measure of publication activity ..	W Glänzel	2006	99
☐	The h index and career assessment by numbers	CD Kelly, MD Jennions	2006	96
☐	On the opportunities and limitations of the H-index	W Glänzel	2006	66
☐	Does the h index have predictive power?	JE Hirsch	2007	102
☐	The R-and AR-indices: Complementing the h-index	B Jin, L Liang, R Rousseau, L	2007	96
☐	Using the H-index of concentration with published data	R Schmalensee	1977	94
☐	... than the h index? A comparison of nine different variants of the h index ...	L Bornmann, R Mutz, HD Dani	2008	52
☐	H-index: an evaluation indicator proposed by scientist	BH Jin	2006	55
☐	Exploring the h-index at the author and journal levels using bibliometric data of ...	G Saad	2006	53

Figure 1. Search results for “h-index“ based on a Google Scholar search. (Date of search: 04/15/2010)

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 freely available for everyone and can be used without fees.

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 here, 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

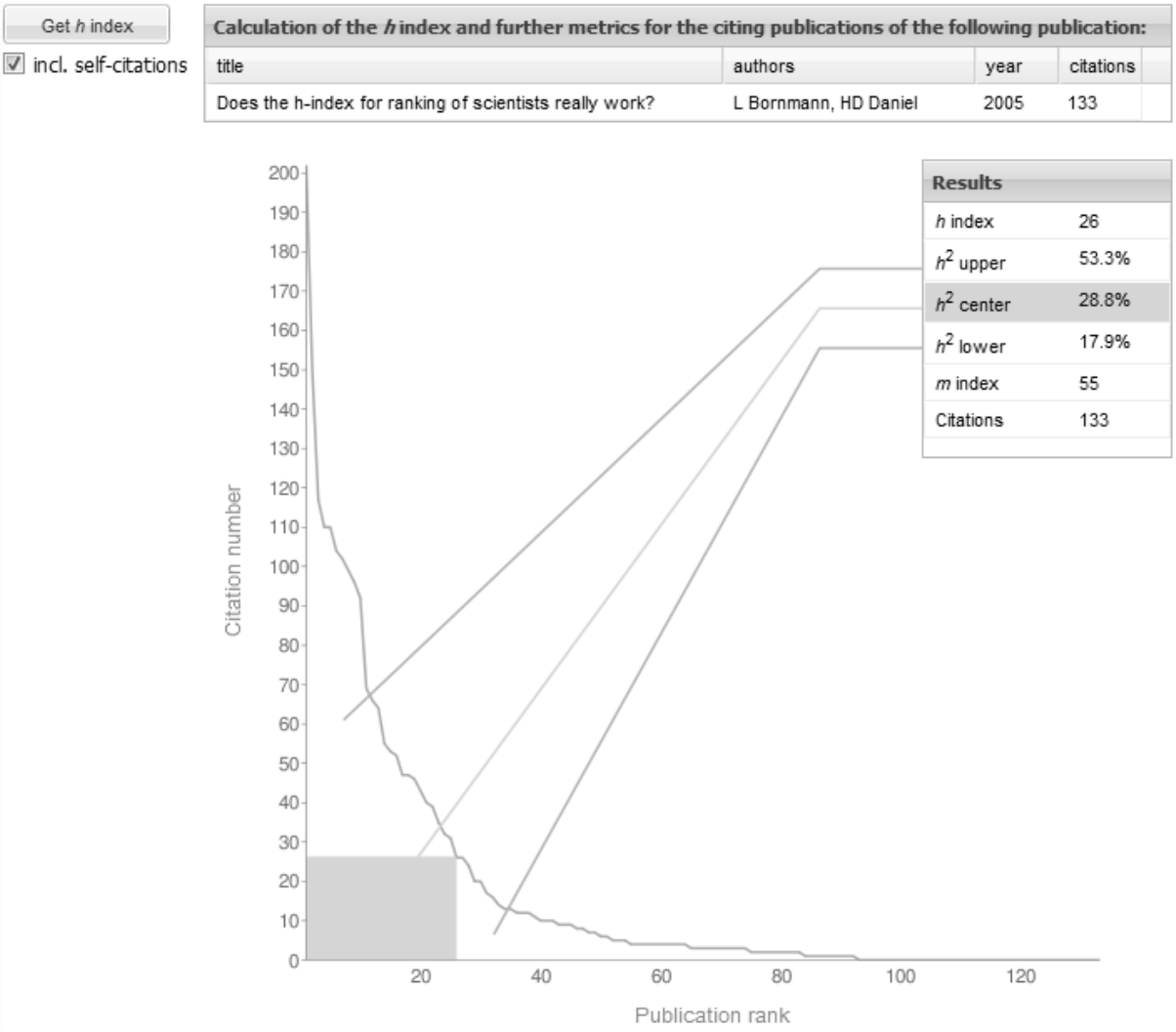


Figure 2. Analysis result for the publication "Does the *h*-index for ranking of scientists really work" by Bornmann and Daniel (2005) including self-citations. The result comprises *h* index, h^2 upper, h^2 center, h^2 lower, *m* index, and the total number of citations. Additionally, the citation distribution of the rank-ordered citing publications is plotted. (Date of analysis: 04/15/2010)

in GS brings about 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 center, h^2 upper, the single publication *m* index, and the total citation counts (#citations) are displayed (see Figure 2). The bottom part of the Web page 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

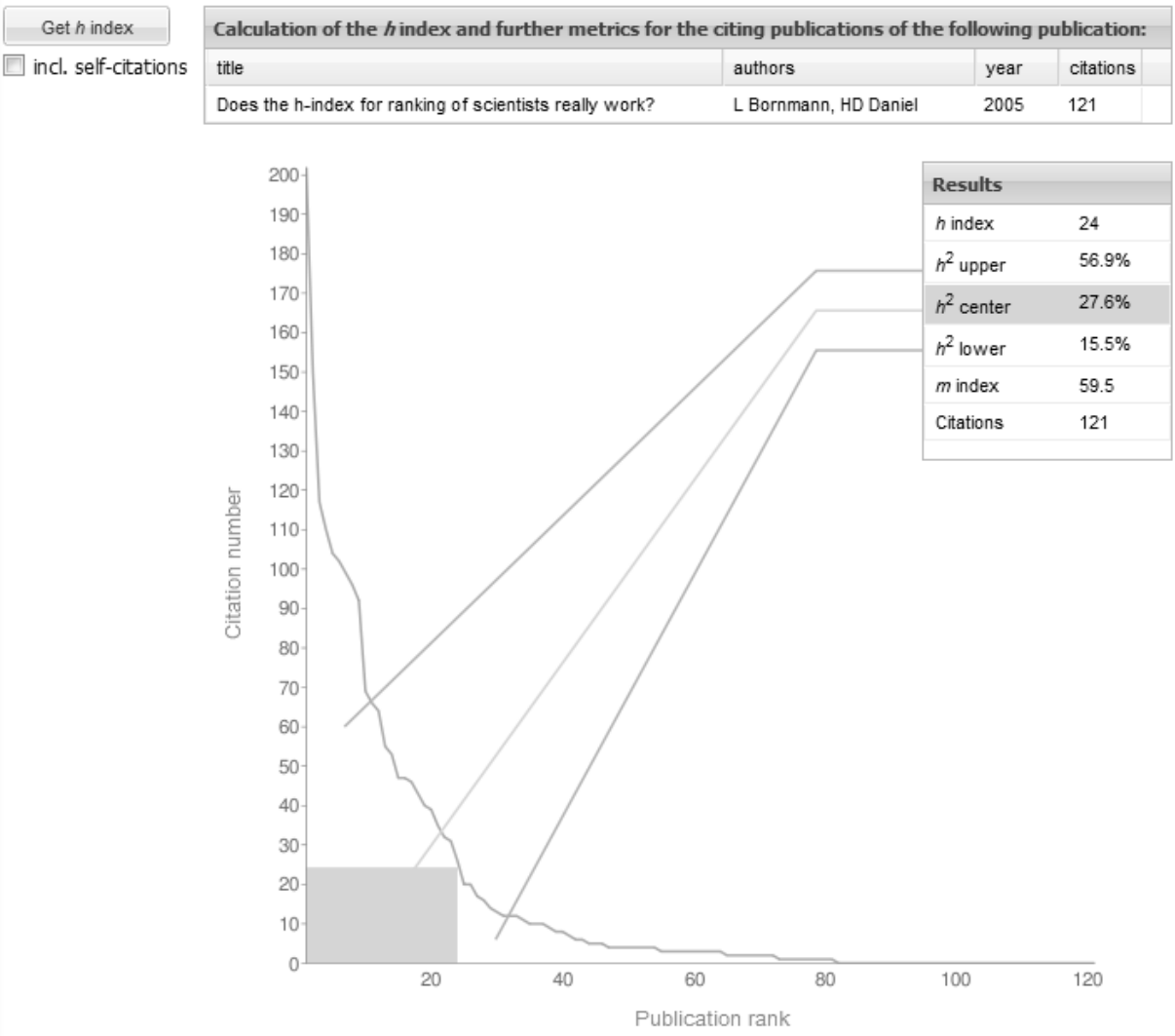


Figure 3. Analysis result for the publication "Does the h-index for ranking of scientists really work" by Bornmann and Daniel (2005) excluding self-citations. The elimination of self-citations leads to slightly different results (see Figure 2 for comparison). (Date of analysis: 04/15/2010)

for each citing publication. In addition, the graph visualizes the area proportions h^2 lower, h^2 center, and h^2 upper. With a percentage of 53.3% for h^2 upper it is clearly visible 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 thereto 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

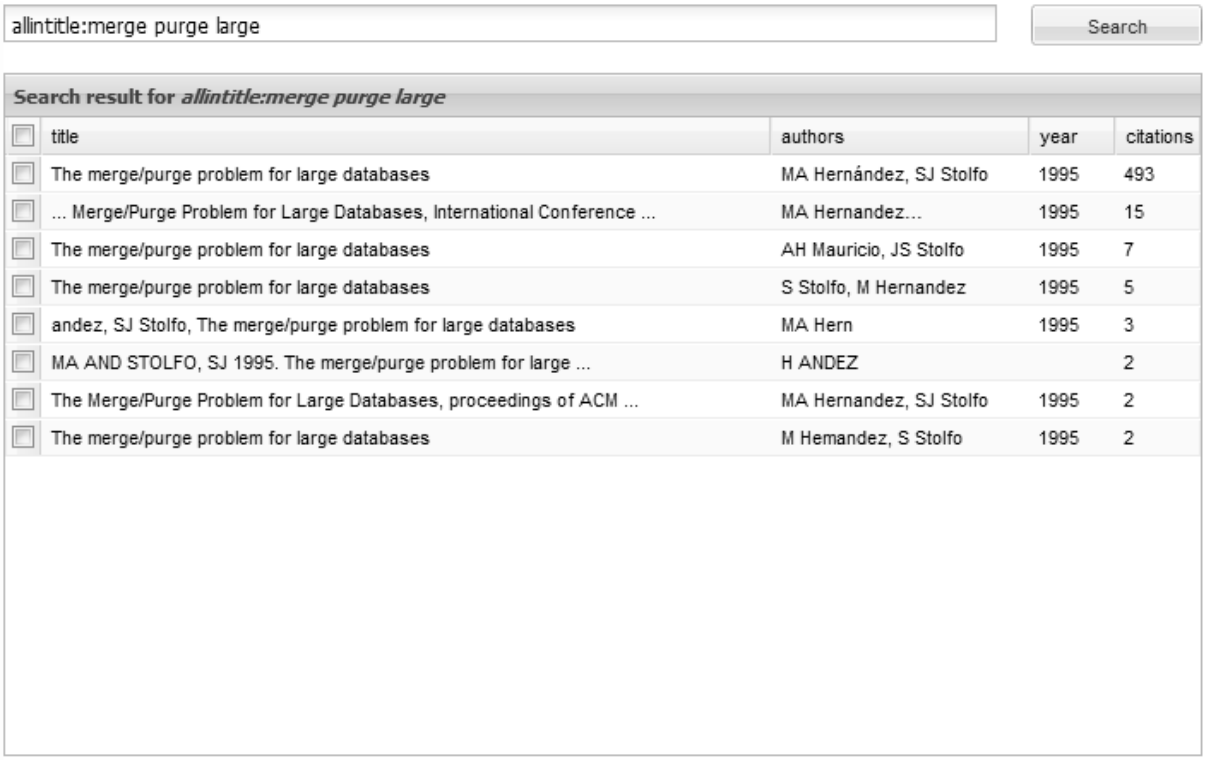


Figure 4. Search result for "allintitle:merge purge large". All retrieved Google Scholar records refer to the same publication by Hernández and Stolfo (1995). (Date of search: 04/15/2010)

publication are filtered away for the calculation of the performance indicators. In the example of Figure 3, the number of citations decreases 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 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 data quality problems of GS. In particular, GS often contains duplicates for the same real world publication. For example, Figure 4 shows a search result for "allintitle:merge purge large" consisting of eight GS records that all refer to the same real-world publication (that ironically deals with the automatic identification of duplicates) (Hernández & Stolfo, 1995). The user of our Web application may therefore mark all duplicates of a publication in question and then perform a joint citation analysis. To this end, the Web application retrieves the lists of citing publication for each selected GS record individually and subsequently merges them into a combined list. In doing so, the Web

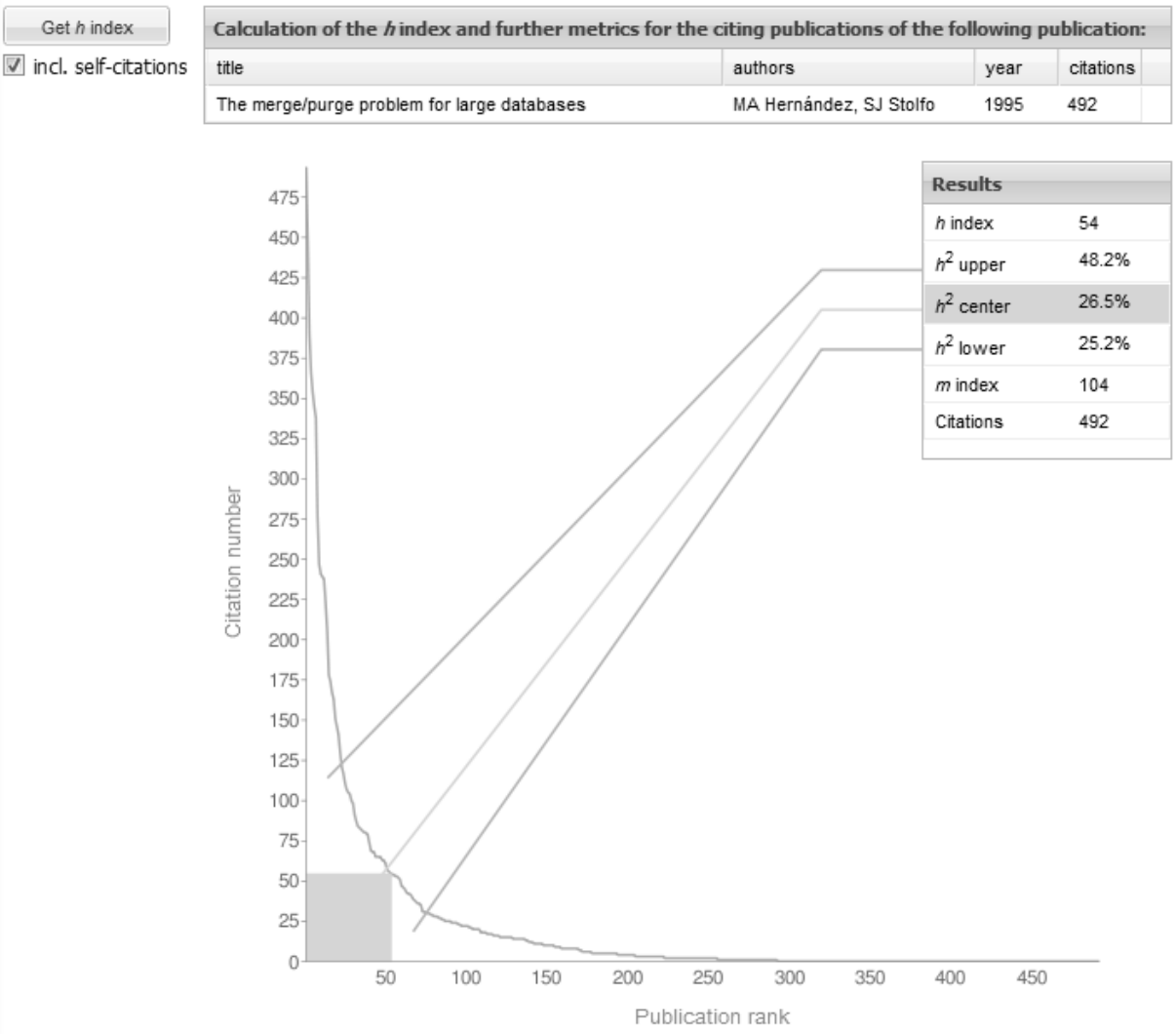


Figure 5. Analysis result for the top record of the search displayed in Figure 4. (Date of analysis: 04/15/2010)

application ensures that all relevant citations are taken into account even if they refer to different GS records. The consideration 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 duplicates leads to a higher number of citations (528 vs. 492) and to a slightly higher *h* index value (55 vs. 54).

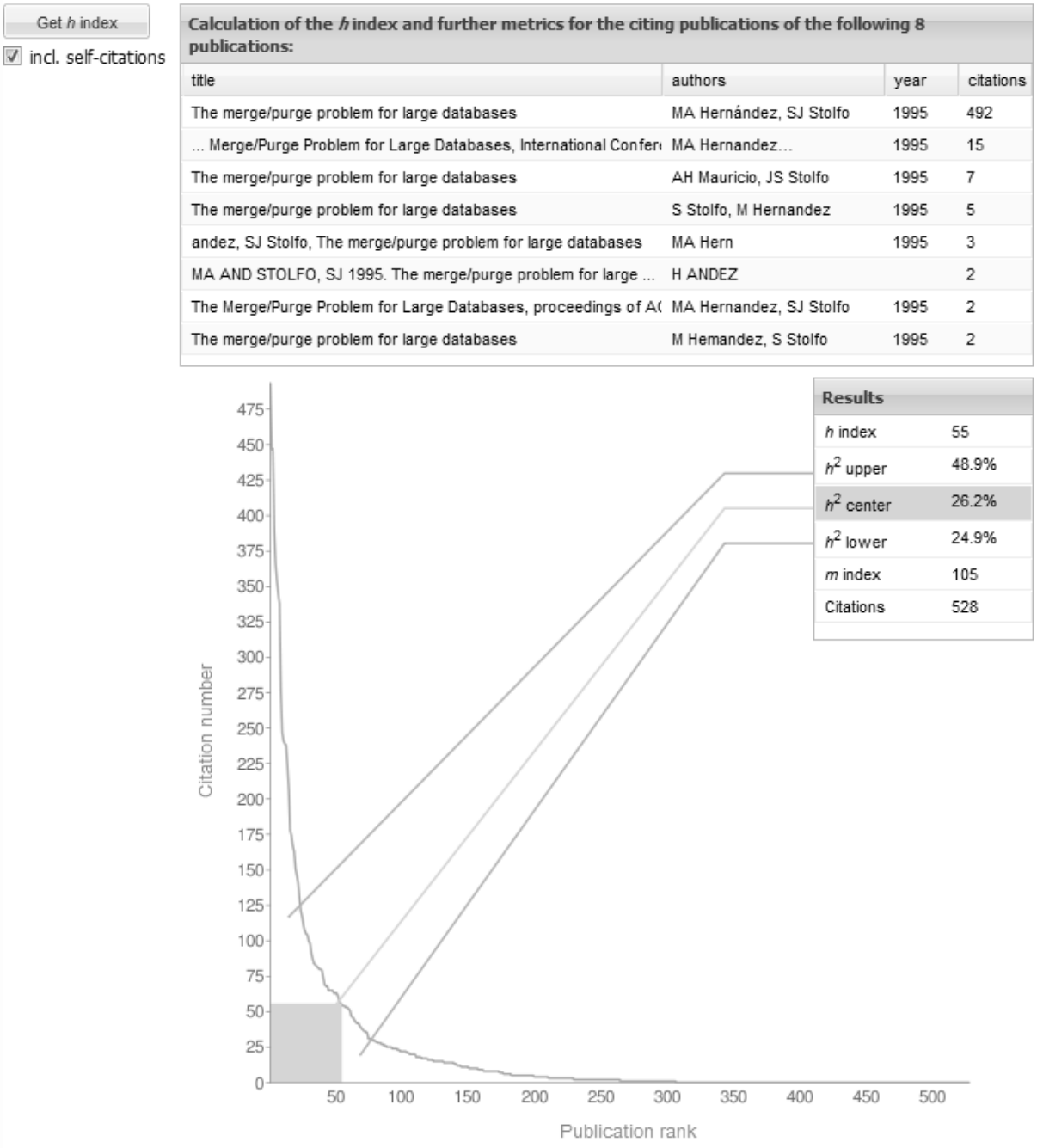


Figure 6. Joint analysis result for all publication records shown in Figure 4. The consideration of all duplicates of the same real-world publication influences the results of the analysis, e.g., the *h* index. (Date of analysis: 04/15/2010)

Discussion

GS has been available on the Internet since 2004 as 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 & Daniel, 2008). According to Harzing and van der Wal (2008)

"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, PA, USA) 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" (p. 62). In addition, GS does not search only peer-reviewed and professional research journals (as WoS does, for example): "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 for publications that are not published in journals indexed by Thomson Reuters (or in Scopus, provided by Elsevier, Amsterdam, the Netherlands), 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 chapter).

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 Jacso (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). Jacso (2010) reviews the recent developments in Google Scholar's management of bibliographic metadata and acknowledges the usefulness of GS's keyword search but, on the other hand, 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 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" (Jacso, 2008b, p. 270). Upon the background of 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 improvable GS data quality lies in the automatic generation of the GS data set (among other things, automatic extraction of references 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 the journals or conferences) and errors in the metadata (e.g., due to typographical errors in titles, extraction errors when splitting reference strings). The performances measures calculated by our Web application may not be influenced significantly by the questionable data quality as other measures, like 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 which relies on the publication year of the citing publications. However, missing citations and phantom citations in GS, of course, may 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 real world publication. The identification and summary 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 & Rahm, 2007). At present, 100% correct automatic identification of duplicates is not possible. Therefore, the

user of our Web application has to select manually all duplicates before calculating the performance measures.

All in all, the citations performances 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 proofs – whenever possible – the convergent validity (Bornmann et al., 2009) 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, 74*(2), 257-271.

Bornmann, L., & Daniel, H.-D. (2007). What do we know about the *h* index? *Journal of the American Society for Information Science and Technology, 58*(9), 1381-1385.

Bornmann, L., & Daniel, H.-D. (2009). The state of *h* index research. Is the *h* index the ideal way to measure research performance? *EMBO Reports, 10*(1), 2-6.

Bornmann, L., Marx, W., Schier, H., Rahm, E., Thor, A., & Daniel, H. D. (2009). 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, 3*(1), 27-35. doi: DOI 10.1016/j.joi.2008.11.001.

Bornmann, L., Mutz, R., & 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, 59*(5), 830-837.

Bornmann, L., Mutz, R., & Daniel, H.-D. (2010). The *h* index research output measurement: two approaches to enhance its accuracy. *Journal of Informetrics, 4*(3), 407-414.

Bornmann, L., Mutz, R., Daniel, H.-D., Wallon, G., & Ledin, A. (2009). Are there really two types of *h* index variants? A validation study by using molecular life sciences data. *Research Evaluation, 18*(3), 185-190.

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

Gardner, S., & Eng, S. (2005). Gaga over Google? Scholar in the social sciences. *Library Hi Tech News, 22*(8).

Giles, J. (2005). Science in the web age: start your engines. *Nature, 438*(7068), 554-555.

Harzing, A. W., & van der Wal, R. (2008). Google Scholar as a new source for citation analysis. *Ethics in Science and Environmental Politics, 8*, 61-73.

Hernández, M. A., & Stolfo, S. J. (1995). The merge/purge problem for large databases. In M. J. Carey & D. A. Schneider (Eds.), *Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data, , May 22-25* (pp. 127-138). San Jose, CA, USA: ACM Press.

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, 102*(46), 16569-16572.

Jacso, P. (2008a). Google Scholar and The Scientist. Retrieved June 5, 2008, from http://www2.hawaii.edu/~jacso/extra/gs/

Jacso, P. (2008b). The plausibility of computing the h-index of scholarly productivity and impact using reference-enhanced databases. *Online Information Review, 32*(2), 266-283. doi: Doi 10.1108/14694520810879872.

Jacso, P. (2010). Metadata mega mess in Google Scholar. *Online Information Review, 34*(1), 175-191. doi: Doi 10.1108/14684521011024191.

Neuhaus, C., & Daniel, H.-D. (2008). Data sources for performing citation analysis - an overview. *Journal of Documentation, 64*(2), 193-210.

Perkel, J. M. (2005). The future of citation analysis. *The Scientist, 19*(20), 24.

Schubert, A. (2009). Using the h-index for assessing single publications. *Scientometrics, 78*(3), 559-565.

Thor, A., & Rahm, E. (2007). MOMA - a Mapping-based Object Matching System. In CIDR (Ed.), *Third Biennial Conference on Innovative Data Systems Research* (pp. 247-258). Asilomar, CA, USA: www.cidrdb.org.

van Eck, N. J., & Waltman, L. (2008). Generalizing the *h*- and *g*-indices. *Journal of Informetrics, 2*(4), 263-271.