The *hw*-rank: An h-index variant for ranking web pages

Judit Bar-Ilan
Department of Information Science
Bar-Ilan University, Israel
email: Judit.Bar-Ilan@biu.ac.il
phone: 972-35318351; fax: 972-37384027

and

Mark Levene
Department of Computer Science and Information Systems
Birkbeck University of London, UK
email: M.Levene@dcs.bbk.ac.uk

Abstract

We introduce a novel ranking of search results based on a variant of the h-index for directed information networks such as the Web. The h-index was originally introduced to measure an individual researcher's scientific output and influence, but here a variant of it is applied to assess the "importance" of web pages. Like PageRank, the "importance" of a page is defined by the "importance" of the pages linking to it. However, unlike the computation of PageRank which involves the whole web graph, computing the h-index for web pages (the *hw*-rank) is based on a local computation and only the neighbors of the neighbors of the given node are considered. Preliminary results show a strong correlation between ranking with the *hw*-rank and Pagerank, and moreover its computation is simpler and less complex than computation of the PageRank. Further, larger scale experiments are needed in order to assess the applicability of the method.

Keywords: hw-rank; h-index; PageRank; ranking of search results

Introduction

Searching for information is one of the major activities on the Web. Due to the vast amounts of information available, ranking search results is crucial for the working of a search engine. In the pre-Web era, search results were ranked either based on information obtained from the document itself (e.g. term frequency) including some information obtained from the collection as a whole (e.g. inverse document frequency) (Salton & McGill, 1986), or were ordered chronologically (mainly in bibliographic databases). The hypertextual structure of the Web allows taking into account the structural information (i.e. links) as well as the content and its metadata. Early attempts considered simply counting the number of incoming links, known as inlinks (Carrière & Kazman, 1997). It was noticed that inlinks, similarly to citations in the academic world are signals of "impact". For a discussion of citation impact, see for example (Garfield, 1973; Moed, 2005), and for a discussion of the impact of links see (Ingwersen, 1998; Thelwall, 2006). A current page on the Google website about search states: "The underlying assumption is that more important websites are likely to receive more links from other websites." (Google, n.d.). However, simply counting inlinks is not sufficient, since, unlike in the academic world, web pages can be set up easily and links are inserted without undergoing a reviewing process, making it is easy to promote a given web page by simply setting up a large number of pages that link to it; in fact such practice is considered by web search engines to be link spam (Gyöngyi & Garcia-Molina, 2005). Thus more complex methods were needed to take into account the hypertextual structure of the Web for search engine result ranking. The best-known methods are PageRank (Brin & Page, 1998; Page, Brin, Motwani & Winograd, 1999) for ranking web pages in a network, and HITS (Kleinberg, 1999) for ranking web pages returned from a user query. Another Google page from 2010 (now non-existing, but retrieved from the Internet Archive) explains: "PageRank interprets a link from Page A to Page B as a vote for Page B by Page A. PageRank then assesses a page's importance by the number of votes it receives. PageRank also considers the importance of each page that casts a vote, as votes from some pages are considered to have greater value, thus giving the linked page greater value" (Google, 2010).

The idea that links from "more important" web pages should count more was not new, and was already suggested for citation networks (Pinski & Narin, 1976) and also for sociometric analysis (Katz, 1953). At that time the methods suggested by Pinski and Narin were not applied widely, and bibliometrics continued to rely mostly on simple citation counts. However, more recently these ideas were revived, probably as a result of the popularity of the PageRank. PageRank type metrics for journals include the SJR (SCImago, 2007; Guerrero-Bote & Moya-Anegón, 2012), the Eigenfactor and the Article Influence (West, Bergstrom & Bergstrom, 2010; eigenfactor.org, 2008). Thus we clearly see the mutual influence of bibliometrics and information retrieval.

The *h-index* is, relatively speaking, a new comer in bibliometrics. It was introduced in 2005 by Hirsch (2005). Originally it was intended to measure the individual researcher's scientific output, a measure that jointly considers publication and citation counts. "A scientist has index h, if h of his or her N_p papers have at least h citations each and the other $(N_p - h)$ papers have $\leq h$ citations each." (Hirsch, 2005, p. 16569). The proposed index was quickly picked up by bibliometricians, who discussed the limitations (e.g. Glänzel, 2006; Costas & Bordons, 2007; Bornmann & Daniel, 2009), suggested variants (e.g. Egghe, 2006; Ruane & Tol, 2008; Guns & Rousseau, 2009) and applied the measure to other datasets, not only to the individual's list of publications (e.g. Braun, Glänzel & Schubert, 2006; van Raan, 2006; Bar-Ilan, 2010a). There is also interest in applying the h-index to graphs, e.g. (Zhao, Rousseau & Ye, 2011; Korn, Schubert & Telcs, 2009).

In this paper we suggest an application of the h-index for ranking web pages, where we consider inlinks on the Web as analogues of citations in scholarly publications.

Ranking with the *hw*-index

The idea for ranking with the *hw*-index is based on Schubert's (2009) extension of the h-index for assessing single publications, and on the lobby index introduced by Korn, Schubert and Telcs (2009). The h-index for assessing single publications assesses the indirect citation influence of the given publication, *p* by considering the number of citations received by the publications citing *p*.

Schubert (2009) defines: "the h-index, h, of a publication as the citation h-index of the set of papers citing it, i.e., not more than h of the papers citing it should receive not less than h citations" (p. 560). More formally, the h-index, *h(p)* of a publication *p* is defined as:

$$h(p) = \max_{h} \text{ there exist } h \text{ citing papers of } p \text{ that received } h \text{ citations or more}$$

We can apply a similar definition to web pages, and assess their importance not by the number of inlinks they receive but by the number of inlinks the pages linking to it receive. More precisely, the *hw*-index, *hw(wp)* of a web page *wp* is defined as:

$$hw(wp) = \max_{h} \text{ there exist } h \text{ web pages linking to } wp \text{ that received } h \text{ inlinks or more}$$

This definition is similar to the definition of the *hw*-index of a web site introduced in (Bar-Ilan, 2010b), and we note that the lobby index (Korn et al., 2009) captures the same idea for undirected networks. The suggested *hw*-index measures the indirect influence of a web page. The basic idea is similar to that of the PageRank: a page is "important" if many "important" pages link to it, but the computation is much simpler and faster; it is not based on the whole web graph but only on the neighbors of the neighbors of the given page, where *a* is a neighbor of *b* if there is a link from *a* to *b*. PageRank involves an eigenvalue computation which is much more expensive. One possible objection to this measure could be that the *hw*-index has only integer values, and thus many pages might receive the same rank. This problem can be partially overcome by computing the hw_{rat} index of each neighboring (inlinking) web page (an analogue of the h_{rat} index (Guns & Rousseau, 2009) and the h^{Δ} index (Ruane & Tol, 2008)), where the hw_{rat} of *wp* is determined as follows, let *n* be the minimum number of additional inlinks that the neighbors of *wp* should receive in order to increase *hw(wp)* by 1, then

$$\begin{aligned} hw_{rat}(wp) &= hw(wp) + \frac{n}{2 * hw(wp) + 1}, \text{ if } hw(wp) < \#\, inlinks\ to\ wp; \\ &= hw(wp), \text{ if } hw(wp) = \#\, inlinks\ to\ wp \end{aligned}$$

The reason that *n* is divided by 2**hw(wp)*+1 is that 2**hw(wp)*+1 is the largest possible increment needed for increasing the *hw*-index from *h* to *h*+1. It should be noted that the maximum value *hw* and hw_{rat} can attain is the number of inlinks of *wp*. hw_{rat} has more discriminating power than the *hw*.

Demonstrating the computation

Unfortunately comprehensive backlink data to web pages obtained through search engines are not readily available any more. In the past Yahoo's Site Explorer used to provide such data, and Google's *link:* modifier, also a source for backlinks, currently displays only an unknown fraction of links pointing to a given page. Thus in this demonstration we rely on data collected in 2010 on pages linking to two pages in Peter Ingwersen's website as of 2010: www.db.dk/pi (and www.db.dk/pi/iri. Archived versions of these pages can be found on the Internet Archive[1]. The first page was Peter Ingwersen's homepage at the Royal School of Library and Information Science, while the second page included information on the book "Information Retrieval Interaction" by Peter Ingwersen. The electronic version of the book was available for downloading from this page. Data were collected on May 8, 2010 (Bar-Ilan, 2010b). At that time Yahoo's Site Explorer identified 127 links to www.db.dk/pi and 83 links to www.db.dk/pi/iri, with 17 pages linking to both target pages. For each of the 193 unique pages that link to either of the two target pages, the number of pages linking to them was recorded, again using the Site Explorer. This allows us to compute both *hw* and hw_{rat} for both pages. The *hw* index for the home page was 22, since there were 22 pages which received at least 22 links each, the number of inlinks of the linking pages ranged between 908 and 25. The 23rd rank ordered page received 22 links, thus the hw_{rat} for the home page is 22+23/45=22.511. Similarly the *hw* index for the book page was 17, with the 17th ranked ordered page receiving 19 links and 18th ranked ordered page receiving 16 links, thus the hw_{rat} of this page is 17+17/35=17.514. The rank ordered linking pages for both pages are plotted in Figures 1 and 2.

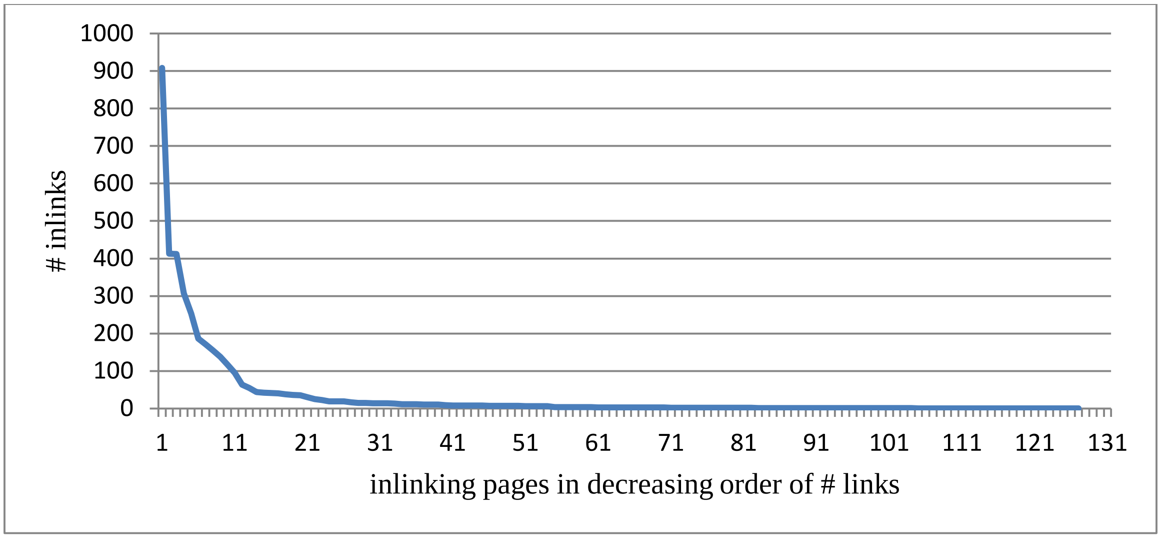

Fig 1. Inlink distributions of pages linking to www.db.dk/pi

[1] https://web.archive.org/web/20100401164532/http://www.db.dk/ombiblioteksskolen/medarbejdere/default.asp?cid=684&tid=4 and https://web.archive.org/web/20091125213609/http://vip.db.dk/pi/iri/index.htm

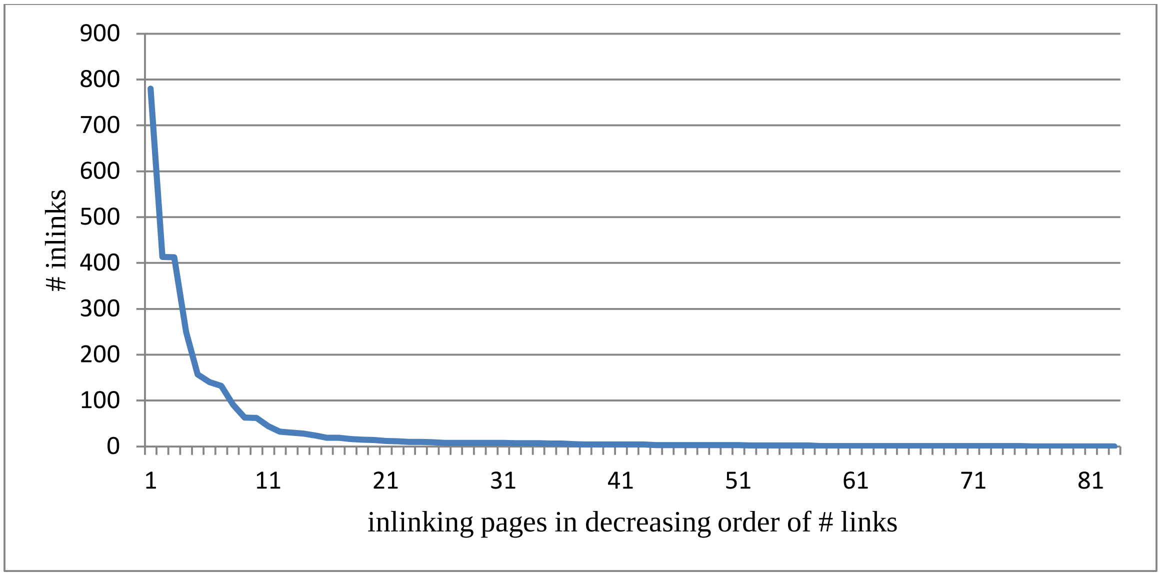

Fig 2. Inlink distributions of pages linking to www.db.dk/pi/iri

Preliminary results

We experimented with this web graph available at http://snap.stanford.edu/data/web-Google.html. The web graph consists of about 875,000 nodes and 5.1 million edges. Both the PageRank and *hw*-rank were computed for the nodes of this graph. The Spearman correlation on the top-1000 results was 0.798 (p<.001). We also ran the same analysis on a graph that represents all Wikipedia administrator elections. In order for a Wikipedia contributor to become an administrator (users with additional rights) the Wikipedia community via a public discussion or a vote decides who to promote to an admin user. The nodes of the graph we analyzed represent the users and the links represent one user voting for another one. The data can be found at http://snap.stanford.edu/data/wiki-Vote.html. This graph consists of about 7,100 nodes and 100,000 edges. For this graph the Spearman correlation between the PageRank-ranked and the *hw*-ranked results for the top-1000 results was 0.983 (p<.001). We note that in both cases the correlation is strong, which indicates a monotonic relationship between the PageRank and the *hw*-rank. However, the relationship is not necessarily linear (Hauke & Kossowski, 2011), and thus more research needs to be done to establish any concrete connections between the two ranking measures. (For the comparison with PageRank, we did not use an existing library, but rather the PageRank algorithm was implemented from scratch and optimised for the purpose of the research.) It is not surprising that the rankings based on the *hw*-index and PageRank are strongly correlated, since they both measure the authority of a web page through linkage. It should be noted that the ranking based on the number of inlinks is also strongly correlated with PageRank (Upstill, Craswell & Hawking, 2003, Fortunato, Boguňá, Flammini, & Menczer, 2008); however inlink counts are much more susceptible to link spam than the h-index.

Summary and future directions

The primary aim of this short paper was to show that information retrieval can be informed by bibliometrics, and that better interaction between the two communities can lead to interesting complementary developments and possible algorithmic improvements.

The preliminary results are promising, however further extensive studies are needed to decide on the applicability of this measure, including extensive user studies to compare rankings based on the *hw*-index with rankings based on the PageRank. The computation of the *hw*-index is simpler and thus much more efficient than that of PageRank, since it involves only a local computation of two levels from any web page, rather than a computation on the whole web graph, The *hw*-index computation thus scales linearly in the number of web pages in the graph, and can be recomputed locally as the web graph evolves. On the other hand the *hw*-index might be more susceptible to link spam than PageRank, since only the second-order neighborhood of a node is involved in the computation.

A further research direction we are currently exploring is the application of the h-index for measuring popularity of queries over time, as popularity is a metric used by search engines in the ranking of web pages. For this we introduce *m-popularity* and the *m-index*.
A query is *m-popular* if for at least *m* time points at a given granularity, its popularity was greater or equal to *m;* the *m-index* for a query is thus the largest *m* for which the query is *m-popular*. *m-popularity* can be made more robust by, for example, multiplying it by some popularity threshold, say *T*. This would imply that an *m*-popular query would have passed the threshold by *mT* for at least *m* time points. Another variation would, for example require that the *m* time points are consecutive or temporally *close* to each other in some precise sense. One application of *m*-popularity is that it would enable a search engine to distinguish between queries which are only popular for short periods as opposed to ones that are popular over a long time span.

Acknowledgement

The authors thank Chatzimichalis Konstantinos for carrying out the comparisons between the *hw*-index and PageRank while studying at Birkbeck University of London.

References

Bar-Ilan, J. (2010a). Rankings of information and library science journals by JIF and by h-type indices. *Journal of Informetrics*, 4, 141-147.

Bar-Ilan, J. (2010b). The WIF of Peter Ingwersen's website. In Birger Larsen, Jesper Wiborg Schneider and Fredrik Åström (Eds.) *The Janus Faced Scholar: A Festschrift in Honour of Peter Ingwersen*, pp. 119-125. Retrieved from http://lup.lub.lu.se/luur/download?func=downloadFile&recordOId=1632623&fileOId=1632624

Bornmann, L., & Daniel, H.-D. (2009). The state of h-index research. *EMBO reports*, 10 (1), 2-6.

Braun, T., Glänzel, W., & Schubert, A. (2006). A Hirsch-type index for journals. *Scientometrics*, 69(1), 169-173.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30, 107-117.

Carrière, S. J., & Kazman, R. (1997). WebQuery: Searching and visualizing the Web through connectivity. *Computer Networks and ISDN Systems*, 29 (8-13), 1257-1267.

Costas, R., & Bordons, M. (2007). The h-index: Advantages, limitations and its relation with other bibliometric indicators at the micro level. *Journal of Informetrics*, 1(3), 193-203.

Egghe, L. (2006). Theory and practise of the g-index. *Scientometrics*, 69(1), 131-152.

eigenfactor.org (2008). Eigenfactor^{TM} score and Article Influence^{TM} score: Detailed methods. Retrieved from http://www.eigenfactor.org/methods.pdf

Fortunato, S., Boguňá, M., Flammini, A., & Menczer, F. (2008). Approximating PageRank from in-degree. Algorithms and Models for the Web Graph. WAW 2006. *Lecture Notes in Computer Science*, 4936, 59-71.

Garfield, E. (1973). Citation frequency as a measure of research activity and performance. *Essays of an Information Scientist*, vol. 1, 406-406.

Glänzel, W. (2006). On the opportunities and limitations of the h-index. Science Focus, 1(1), 10-11. English version retrieved from http://eprints.rclis.org/9378/1/H_Index_opprtunities.pdf

Gyöngyi, Z, & Garcia-Molina, H. (2005). Link spam alliances. In *Proceedings of the 31st International Conference of Very Large Databases (VLDB)*, pp. 517-528.

Google. (2010). *Corporate information – Technology overview*. Retrieved from https://web.archive.org/web/20100419191933/http://www.google.com/intl/en_uk/corporate/tech.html

Google (n. d.). *Facts about Google and competition*. Retrieved from http://www.google.com/competition/howgooglesearchworks.html

Guerrero-Bote, V. P., & Moya-Anegón, F. (2012). A further step forward in measuring journals' scientific prestige: The SJR2 indicator. *Journal of Informetrics*, 6, 674-688.

Guns, R., & Rousseau, R. (2009). Real and rational variants of the h-index and the g-index. *Journal of Informetrics*, 3, 64-71.

Hauke, J. & Kossowski, T. (2011). Comparison of values of Pearson's and Spearman's correlations on the same sets of data. Quaestiones Geographicae, 30(2), 87-93.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences* (PNAS) 102(46), 16569-16572.

Ingwersen, P. (1998). The calculation of web impact factors. *Journal of Documentation*, 54(2), 236-243.

Katz, L. (1953). A new status index derived from sociometric analysis. *Psychometrika*, 18(1), 39-43.

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5), 604-632.

Korn, A., Schubert, A., & Telcs, A. (2009). Lobby index in networks. *Physica A*, 388, 2221-2226.

Moed, H. F. (2005). *Citation analysis in research evaluation*. Springer, Dortrecht, Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). The PageRank citation ranking: Bringing order to the Web. *Technical Report, Stanford InfoLab*. Retrieved from http://ilpubs.stanford.edu:8090/422/

Pinski, G. & Narin, F. (1976). Citation influence for journal aggregates of scientific publications: Theory, with application to the literature of physics. *Information Processing and Management*, 12, 297-312.

Ruane, F., & Tol, R. S. J. (2008). Rational (successive) h-indices: An application to economics in the Republic of Ireland. *Scientometrics*, 75(2), 395-405.

Salton, G., & McGill, M. J. (1986). *Introduction to modern information retrieval*. McGraw Hill, New York.

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

SCImago (2007). *SJR – SCImago journal & country rank*. Retrieved from http://www.scimagojr.com

Thelwall, M. (2006). Interpreting social science link analysis research: A theoretical framework. *Journal of the American Society for Information Science and Technology*, 57(1), 60-68.

Upstill, T., Craswell, N., & Hawking, D. (2003). Predicting fame and fortune: PageRank or indegree? In: *Proceedings of the 8th Australasian Document Computing Symposium*. Retrieved from http://131.107.65.14/pubs/65254/upstill_adcs03.pdf

Van Raan, A. F. J. (2006). Comparison of the Hirsch-index with standard bibliometric indicators and with peer judgment for 147 chemistry research groups. *Scientometrics*, 67(3), 491-502.

West, J. D., Bergstrom, T. C., & Bergstrom, C. T. (2010). The Eigenfactor^{TM} Metrics: A network approach to assessing scholarly journals. *College and Research Libraries*, 71(3): 236-244.

Zhao, S. X., Rousseau, R., & Ye, F. Y. (2011). h-degree as a basic measure in weighted networks. *Journal of Informetrics*, 5, 668-677.