Google Scholar's Ranking Algorithm: The Impact of Articles' Age (An Empirical Study)

Jöran Beel & Bela Gipp
Otto-von-Guericke University
Department of Computer Science
ITI / VLBA-Lab / Scienstein
Magdeburg, Germany
j.beel|b.gipp@scienstein.org

Abstract

Google Scholar is one of the major academic search engines but its ranking algorithm for academic articles is unknown. In recent studies we partly reverse-engineered the algorithm. This paper presents the results of our third study. While the first study provided a broad overview and the second study focused on researching the impact of citation counts, the current study focused on analyzing the correlation of an article's age and its ranking in Google Scholar. In other words, it was analyzed if older/recent published articles are more/less likely to appear in a top position in Google Scholar's result lists. For our study, age and rankings of 1,099,749 articles retrieved via 2,100 search queries were analyzed. The analysis revealed that an article's age seems to play no significant role in Google Scholar's ranking algorithm. It is also discussed why this might lead to a suboptimal ranking.

1. Introduction

With increasing use of academic search engines it becomes increasingly important for scientific authors that their research articles are well ranked in those search engines in order to reach their audience. To optimize research papers for academic search engines, such as Google Scholar or Scienstein.org, knowledge about ranking algorithms is essential. For instance, if search engines consider how often a search term occurs in an article's full text, authors should use the most relevant keywords in their articles whenever possible to achieve a top ranking.

For users of academic search engines, knowledge about applied ranking algorithms is also essential for two reasons. Firstly, users should know about the algorithms in order to estimate the search engine's robustness to manipulation attempts by authors and spammers and therefore the trustworthiness of the results. Secondly, knowledge of ranking algorithms enables researchers to estimate the usefulness of results in respect to their search intention. For instance, researchers interested in the latest trends should use a search engine putting high weight on the publications' date. Users searching for standard literature should choose a search engine putting high weight on citation counts. In contrast, if a user searches for articles from authors advancing a view different from the majority, search engines putting high weight on citation counts might not be appropriate.

Therefore, this paper deals with the question of how Google Scholar ranks its results. The paper is structured as follows. In the second section related work about Google Scholar's ranking algorithm is presented. The third section covers the research objectives while the fourth section explains the utilized methodology. Finally, the results and their interpretation follow.

2. Related Work

Due to different user needs, many academic databases and search engines enable the user to choose a ranking algorithm. For instance, *ScienceDirect* lets users select between date and relevance[1], *IEEE Xplore* offers in addition a ranking by title and *ACM Digital*

[1] 'Relevance' in most cases means that the more often a search term occurs in a document, the more relevant it is considered.

Library allows users to choose whether to sort results by relevance, publication date, alphabetically by title or journal, citation counts or downloads. However, these 'algorithms' can be considered trivial since users can select only one ranking criteria and are not allowed to use a (weighed) combination of them.

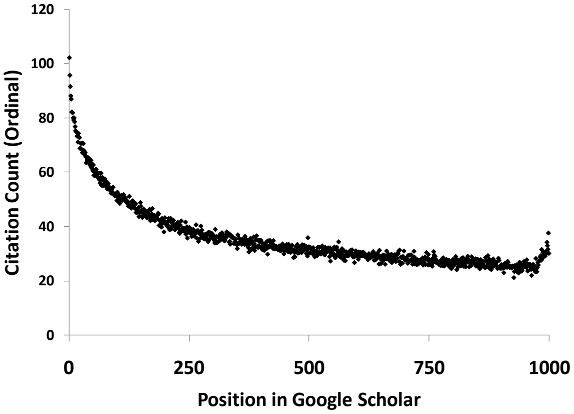

Figure 1: Mean Citation Count

Google Scholar is one of the few academic search engines combining several approaches in a single algorithm[2]. Several studies about Google Scholar exist. For instance, about data overlap with other academic search engines such as Scopus and Web of Science [1], [2], Google Scholar's coverage of the literature in general and in certain research fields [3], [4], the suitability to use Google Scholar's citation counts for calculating bibliometric indices such as the h-index [5] and the reliability of Google Scholar as a serious information source in general [6], [7]. Google Scholar itself publishes only vague information about its ranking algorithm: Google Scholar sorts "articles the way researchers do, weighing the full text of each article, the author, the publication in which the article appears, and how often the piece has been cited in other scholarly literature" [8]. Any other details or further explanation is not available.

Although Google Scholar's ranking algorithm has a significant influence on which academic articles are read by the scientific community, we could not find any studies about Google Scholar's ranking algorithm despite our own ones [9], [10]. From our previous studies we know that

- Google Scholar's ranking algorithm puts high weight on words in the title.
- Google Scholar considers only those words that are directly included in an article and does not consider synonyms of those words.
- Google Scholar seems to put no or low weight on the frequency with which search terms occur in the full text. That means an article will not be ranked higher for a certain search just because the search term occurs frequently in the full text.
- Google Scholar is not indexing text embedded via pictures.
- Google Scholar uses different ranking algorithms for a keyword search in the full text, keyword search in the title, the 'related articles' function and the 'cited by' function.
- Google Scholar's ranking algorithm puts high weight on author and journal names.
- Google Scholar's ranking algorithm weighs heavily on articles' citation counts (see Figure 1), whereas different patterns were discovered.

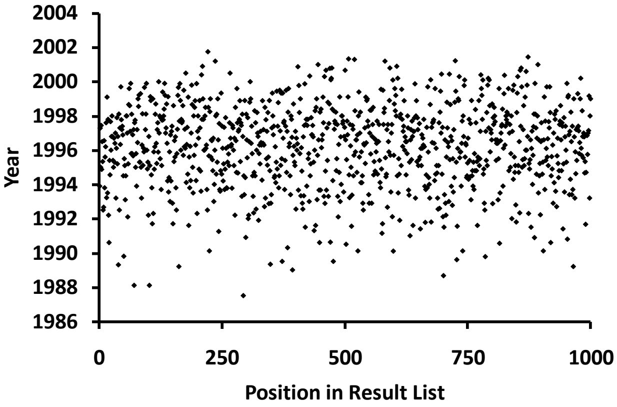

Figure 2: Mean Publication Year per Position

Since citation counts have a strong impact on Google Scholar's rankings, one could assume that older articles are found more often in top positions, since older publications naturally have had more time to be cited. As a consequence, this practice would strengthen the Matthew Effect[3]. To counteract the Matthew Effect and since one might assume that most researchers have an interest in the most recent research results rather

[2] Others are, for instance, *CiteSeer* and *Scienstein.org* [11, 12]

[3] The Matthew Effect describes that well known authors are more often cited just because they are well known [13]. Related to search engines this means: Articles with many citations will be more likely displayed in top positions, therefore get more readers and receive more citations, which then consolidate their lead over lesser cited articles.

than old ones, it seems plausible to rank recent articles better than older ones.

Our previous research indicated that publications from all years are approximately evenly distributed throughout Google Scholars' result list (see Figure 2). Therefore we concluded that an article's age plays a significant role in Google Scholar's ranking algorithm. However, the sample size was small, so further research was needed to confirm or reject this first conclusion.

3. Research Objective

The research objective of the current study was to analyze whether Google Scholar considers articles' age in its ranking algorithm and if so to what extent.

Since Google Scholar offers two search modes (search in title and search in full text) and our previous study indicated that both search modes apply different ranking algorithms we also researched whether Google Scholar's different ranking algorithms weigh differently on an articles' age.

4. Methodology

Google Scholar displays for most articles their publication year in the result list. To obtain publication years for a significant number of papers, we developed a Java program to parse Google Scholar. This program sends search queries to Google Scholar and stores publication years and positions of all returned results in a .csv file. Due to Google Scholar's limitations, only a maximum of 1,000 results per search query was retrievable. The parsing process was performed twice, each time with 1,050 search queries whereas the 1,050 search queries consisted of 350 single-word search queries, 350 double-word search queries and 350 triple-word search queries[4]. In the first run, search terms were searched in the full text. In the second run, search terms were searched in the title.

Table 1: Amount of Search Results by Number of Search Terms (Full Text Search)

		Number of Search Results							
		[0,1]	[2, 10]	[11, 50]	[51, 250]	[251, 1000]	[1001, 10000]	[10001, *]	Total
Single Terms	Absolute	0	0	0	0	0	2	348	350
	Relative	0,0%	0,0%	0,0%	0,0%	0,0%	0,6%	99,4%	100%
Double Term	Absolute	0	0	0	0	3	24	323	350
	Relative	0,0%	0,0%	0,0%	0,0%	0,9%	6,9%	92,3%	100%
Triple Term	Absolute	0	0	0	1	4	86	259	350
	Relative	0,0%	0,0%	0,0%	0,3%	1,1%	24,6%	74,0%	100%
Total	Absolute	0	0	0	1	7	112	930	1050
	Relative	0,0%	0,0%	0,0%	0,1%	0,7%	10,7%	88,6%	100%

From 1,050 full text searches, all search queries returned two or more results (see Table 1) and could be used for the analysis. From 1,050 title searches, 511 returned either a zero or one result and were not considered for further analysis (see Table 2). This was caused by the way search queries were created. They were created automatically by combining different words from a word list which resulted in some senseless search queries such as 'finish father' or 'excessive royalty'. While sufficient documentation exists in which, for instance, the words 'finish' and 'father' occur somewhere in the full text, no documents exist which include these words in the title.

Table 2: Amount of Search Results by Number of Search Terms (Title Search)

		Number of Search Results							
		[0,1]	[2, 10]	[11, 50]	[51, 250]	[251, 1000]	[1001, 10000]	[10001, *]	Total
Single Terms	Absolute	0	1	1	12	23	102	211	350
	Relative	0,0%	0,3%	0,3%	3,4%	6,6%	29,1%	60,3%	100%
Double Term	Absolute	166	89	54	27	11	3	0	350
	Relative	47,4%	25,4%	15,4%	7,7%	3,1%	0,9%	0,0%	100%
Triple Term	Absolute	345	5	0	0	0	0	0	350
	Relative	98,6%	1,4%	0,0%	0,0%	0,0%	0,0%	0,0%	100%
Total	Absolute	511	95	55	39	34	105	211	1050
	Relative	48,7%	9,0%	5,2%	3,7%	3,2%	10,0%	20,1%	100%

Overall, data from 1,561 search queries (1,050 searches in the full text and 511 searches in the title) was used for further analysis. The 1,561 search queries returned a total of 1,364,757 results (1,032,766 articles for full text searches and 331,991 articles for title searches). For 810,793 of the 1,032,766 articles retrieved via full-text search and 288,956 of the 331,991 articles retrieved via title search, Google Scholar displayed the publication year. Those years and the articles' rankings were stored and analyzed. To verify correct execution of the Google Scholar parser, spot checks were performed.

All results of the search queries were visualized as graphs to recognize patterns. In addition, the mean, median, and modal of each position was calculated and displayed in a graph. Overall, a total of 1,567 graphs were created and inspected individually.

5. Results

On first glance, results of the current study seem to confirm our previous results. Graphs of individual search queries show no significant interdependency between an article's age and its ranking in Google Scholar (see also [10]). This is true for all kind of search queries such as searches in full-text or title and searches with single-word, double-word and triple-word queries. The graphs show that publications from all years are evenly distributed throughout the result list (see Figure 3, Figure 4 and Figure 5)[5].

[4] The words for creating the search queries were extracted from an academic word list [14]

[5] The graphs also show that Google Scholar has far more documents from the 90s and current decade in its database

However, looking at the average age, another impression evolves. Figure 6 displays the average publication year (mean) for each position in Google Scholar. It shows clearly that in the top positions articles are on average older than articles in the remaining positions[6].

A look at the numbers confirms this assumption. While those papers ranked in position 1 by Google Scholar were on average published in 1992, papers on position 5 were on average published in 1993, papers on position 100 in 1994 and papers on position 500 in 1995 (see Table 3). Graphs for title-searches look similar (see Figure 7) and no significant differences occurred between single-word, double word and triple word search queries[7].

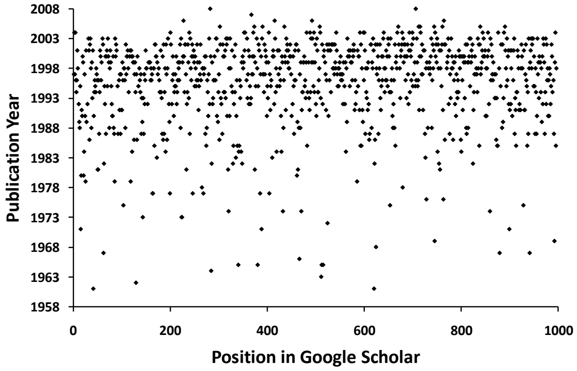

Figure 3: Search Query 'Future'

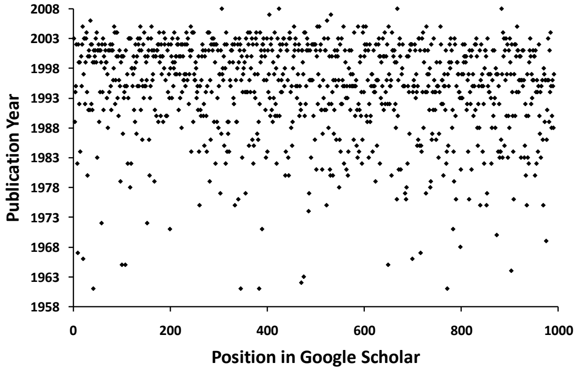

Figure 4: Search Query 'Google Scholar'

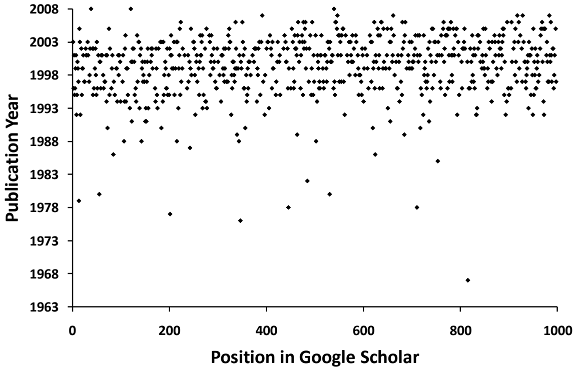

Figure 5: Search Query 'Climate Change Discussion'

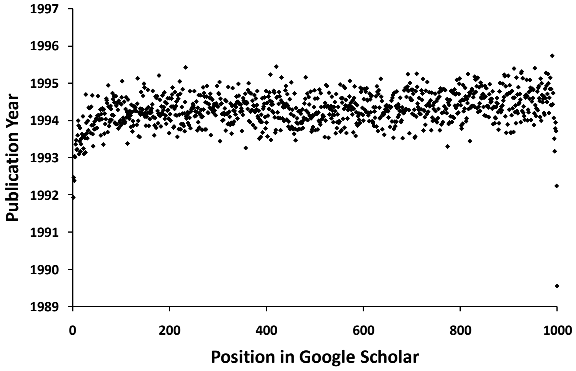

Figure 6: Mean Publication Year (Full-Text Search)

Table 3: Mean Publication Year (Selected Positions)

Position	Publication Year (Mean)	Position	Publication Year (Mean)
1	1992	10	1994
2	1992	50	1994
3	1992	100	1994
4	1993	250	1994
5	1993	500	1995

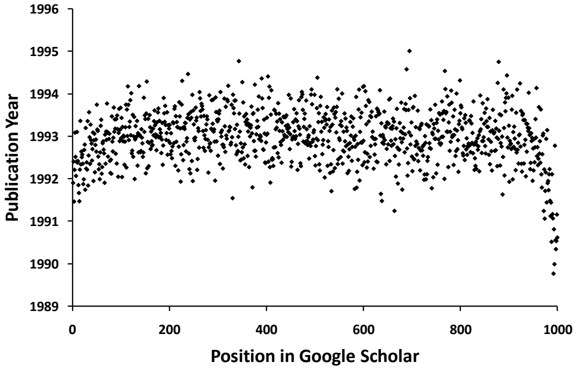

Figure 7: Mean Publication Year (Title Search)

than from decades before. However, this is out of the current study's scope.

[6] Graphs for the modal and median publication year show similar pictures.

[7] In all graphs, some outliers can be observed in the very last positions. This is due to Google Scholar which often does not return the very last results. Therefore the means for the last positions was based on few sample data and hence some outliers could spoil the results.

6. Interpretation and Discussion

Taking into consideration that Google Scholar might become as popular for academic articles as Google is for web pages, we hope to stimulate a discussion with our research into how ranking algorithms of academic search engines should be designed. We believe that users should be able to adjust ranking algorithms to their individual search intension (search for standard literature, search for latest research trends, search for articles by authors advancing a view different from the mainstream, etc.).

If a search engine does not offer this option, as is the case with Google Scholar, users should at least have basic knowledge about the applied ranking algorithm. Only this way they can assess the suitability of an academic search engine for their search intension.

Our research shows that in Google Scholar older articles are found more often in top positions than recent articles. This is probably due to Google Scholar’s strong focus on citation counts and due to Google Scholar putting no or low weight on an article’s publication date. As a consequence, Google Scholar is rather suitable for finding standard literature than the latest research results.

7. Further Research & Data Sharing

This is our third paper about Google Scholar’s ranking algorithm and the algorithm is still far from being known. We invite researchers to join us and would be happy to share our Google Scholar parser and gathered data. Please send us an email if you are interested in the data or software.

8. Acknowledgements

Our thanks go to Ammar Shaker for supporting the development of the Google Scholar parser.

9. References

[1] J. Bailey, C. Zhang, D. Budgen, M. Turner, and S. Charters, “Search engine overlaps : Do they agree or disagree?” in *Second International Workshop on Realising Evidence-Based Software Engineering (REBSE '07)*, 2007, p. 2. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4273274

[2] K. Yang and L. I. Meho, “Citation analysis: A comparison of google scholar, scopus, and web of science,” in *69th Annual Meeting of the American Society for Information Science and Technology*, Austin (US), 2006, pp. 3–8.

[3] W. H. Walters, “Google scholar coverage of a multidisciplinary field,” *Information Processing & Management*, vol. 43, no. 4, pp. 1121–1132, July 2007.

[4] J. J. Meier and T. W. Conkling, “Google scholar’s coverage of the engineering literature: An empirical study,” *The Journal of Academic Librarianship*, vol. 34, no. 34, pp. 196–201, 2008.

[5] J. Bar-Ilan, “Which h-index? - a comparison of wos, scopus and google scholar,” *Scientometrics*, vol. 74, no. 2, pp. 257–271, 2007.

[6] P. Jacso, “Google scholar: the pros and the cons,” *Online Information Review*, vol. 29, no. 2, pp. 208–214, 2005.

[7] B. White, “Examining the claims of google scholar as a serious information source,” *New Zealand Library & Information Management Journal*, vol. 50, no. 1, pp. 11–24, 2006.

[8] (2008) About google scholar. Website. Google Inc. [Online]. Available: http://scholar.google.com/intl/en/scholar/about.html

[9] J. Beel and B. Gipp, “Google scholar's ranking algorithm: An introductive overview (research in progress),” in *Proceedings of 3rd International Conference on Research Challenges in Information Science (RCIS’09)*. IEEE, 2009.

[10] J. Beel and B. Gipp, “Google scholar's ranking algorithm: The impact of citation counts (an empirical study).” to be published, 2009.

[11] B. Gipp and J. Beel, “Scienstein: A research paper recommender system,” in *International Conference on Emerging Trends in Computing*. IEEE, 2009, pp. 309–315.

[12] J. Beel and B. Gipp, “The potential of collaborative document evaluation for science,” in *11th International Conference on Digital Asian Libraries (ICADL'08)*, ser. Lecture Notes in Computer Science (LNCS), G. Buchanan, M. Masoodian, and S. J. Cunningham, Eds., vol. 5362. Heidelberg (Germany): Springer, December 2008, pp. 375–378.

[13] R. K. Merton, “The matthew effect in science,” *Science*, vol. 159, no. 3810, pp. 56–63, January 1968.

[14] S. Haywood. (2008) The academic word list. University of Nottingham. [Online]. Available: http://www.nottingham.ac.uk/ alzsh3/acvocab/wordlists.htm