Google Scholar's Ranking Algorithm: The Impact of Citation Counts (An Empirical Study)

Joeran Beel
Otto-von-Guericke University
Department of Computer Science
ITI / VLBA-Lab / Scienstein
Magdeburg, Germany
j.beel@scienstein.org

Bela Gipp
Otto-von-Guericke University
Department of Computer Science
ITI / VLBA-Lab / Scienstein
Magdeburg, Germany
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 a recent study we partly reverse-engineered the algorithm. This paper presents the results of our second study. While the previous study provided a broad overview, the current study focused on analyzing the correlation of an article's citation count and its ranking in Google Scholar. For this study, citation counts and rankings of 1,364,757 articles were analyzed. Some results of our first study were confirmed: Citation counts is the highest weighed factor in Google Scholar's ranking algorithm. Highly cited articles are found significantly more often in higher positions than articles that are cited less often. Therefore, Google Scholar seems to be more suitable for searching standard literature than for gems or articles by authors advancing a view different from the mainstream. However, interesting exceptions for some search queries occurred. In some cases no correlation existed; in others bizarre patterns were recognizable, suggesting that citation counts sometimes have no impact at all on articles' rankings.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *Information filtering, Search process, Selection process*.

General Terms
Algorithms

Keywords
Academic Search Engines, Google Scholar, Ranking Algorithm, Citation Counts, Empirical Study

1. INTRODUCTION
With the increasing use of academic search engines it becomes increasingly important for scientific authors to have their research articles well ranked in those search engines, in order to reach their audience. In other words, for scientists, knowledge about ranking algorithms is essential in order to optimize their research papers for academic search engines, such as Google Scholar or Scienstein.org. 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 basic reasons. Firstly, users should know about the algorithms in order to estimate the search engine's robustness towards manipulative 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 a high weight on the publications' date. Users searching for standard literature should choose a search engine putting a high weight on citation counts. In contrast, if a user searches for articles by authors advancing a perspective which differs from the majority, search engines putting a high weight on citation counts might not be appropriate.

This paper deals with the question of how Google Scholar ranks its results, and is structured as follows: First, related work is presented. Then, the research objective is outlined, followed by the applied methodology. Finally, results and their interpretations are presented.

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* in addition, offers a ranking by title and *ACM Digital 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.

Ranking academic articles by citation counts is a common procedure, but remains controversial. With regards to academic search engines, two points of criticism are particularly relevant.

Firstly, ranking articles based on citation counts strengthens the Matthew Effect. This means that those articles with many citations are displayed first, therefore they get many readers and receive many citations, which in turn causes them to be displayed first. This is a common problem which exists in the scientific community [1]. However, academic search engines could increase this dilemma as users of search engines usually pay

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

attention only to the first results. A study about web search engines revealed that around 42% of all outgoing clicks were on the result in position 1 [2]. Around 90% of outgoing clicks were on a result on the first page. That means that most users of a web search engine did not even pay attention to the second page. This illustrates the importance for webmasters to be listed in one of the very first positions and that ranking algorithms significantly influences the amount of visitors a webpage receives. It seems likely that the same is basically true for academic search engines.

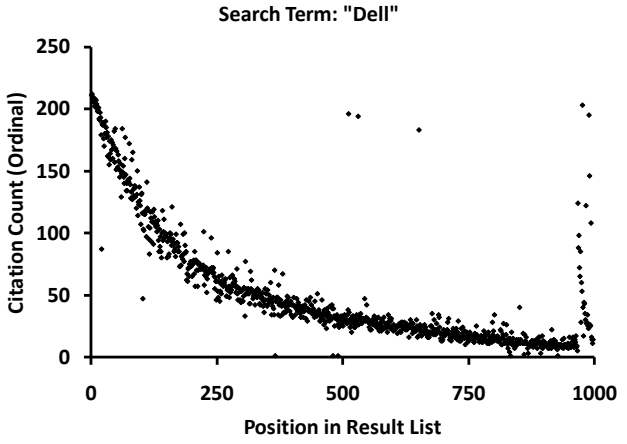

Figure 1: Pattern I

The second point of criticism relates to the fact that citation measures impact but not quality in general [3, 4]. That means, articles with many citation counts are not always 'good'. It might make sense to rank articles with high citation counts first if a user is searching for standard literature with high impact. But there may be situations in which it makes no sense to display highly cited papers in the first positions. This could be, for instance, if someone searches for the latest trends in a certain research field or articles from authors advancing a view different from the majority.

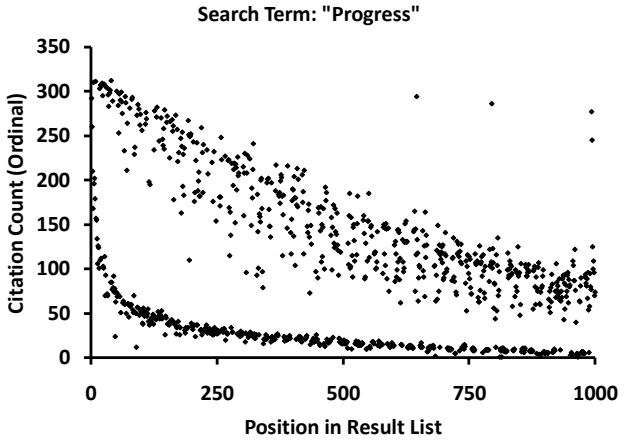

Figure 2: Pattern II

As mentioned, most academic databases offer different approaches for ranking publications and users can select *one* of them. Google Scholar is one of the few academic search engines combining several approaches in a single algorithm[2]. Several studies about Google Scholar exist. Studies include, for instance, research into data overlap with other academic search engines such as Scopus and Web of Science [5], [6], Google Scholar's coverage of the literature in general and in certain research fields [7], [8], the suitability to use Google Scholar's citation counts for calculating bibliometric indices such as the h-index [9] and the reliability of Google Scholar as a serious information source in general [10], [11]. 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" [12]. 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 one [13]. From our previous study we know that:

- Google Scholar's ranking algorithm puts a high weight on words occurring in the title.
- Google Scholar considers only those words that are directly included in an article and does not consider synonyms of those words.
- The frequency in which query terms occur in the full text seems to have little to no impact on Google Scholar's rankings. That means that 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 images.
- 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 a high weight on author and journal names.
- Google Scholar seems to weight recent articles stronger than older articles in order to compensate for the Matthew effect.

The most confusing finding from our previous research was about how Google Scholar weighs citation counts. We found out that in general, Google Scholar weighs an article's citation count heavily. However, different patterns were discovered (see Figure 1 and Figure 2) and we could not explain why these patterns occurred or if further patterns existed.

3. RESEARCH OBJECTIVE

Our previous study indicated a strong interrelationship between an article's citation count and its position in Google Scholar. The research objective of the current study was to confirm or reject the previous results based on a larger sample size and to research whether further patterns exist about how rankings interrelate with citation counts.

Since Google Scholar offers two search modes (search in title and search in full text), and our previous study indicated that different ranking algorithms are applied, we also researched

[2] Others are, for instance, *CiteSeer* and *Scienstein.org* [15, 3]

whether citation counts have a different weight when searching in the title rather than searching the full text.

4. METHODOLOGY

Google Scholar displays for each article its citation count in the results list. To obtain citation counts for a significant number of papers, we developed a Java program to parse Google Scholar[3]. This program sends search queries to Google Scholar and stores the citation counts 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 where 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.

From 1,050 full text searches, all search queries returned more than 50 results (see Table 1) and could be used for the analysis. From 1,050 title searches, 511 returned either 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 1: Amount of Search Results by Number of Query Terms (Full Text Search)

			Number of Search Results						
			[0,1]	[2, 10]	[11, 50]	[51, 250]	[251, 1000]	[1001, 10000]	[10001, *]
Number of Query Terms	Single Words	Absolute	0	0	0	0	0	2	348
		Relative	0.0%	0.0%	0.0%	0.0%	0.0%	0.6%	99.4%
	Double Words	Absolute	0	0	0	0	3	24	323
		Relative	0.0%	0.0%	0.0%	0.0%	0.9%	6.9%	92.3%
	Triple Words	Absolute	0	0	0	1	4	86	259
		Relative	0.0%	0.0%	0.0%	0.3%	1.1%	24.6%	74.0%
	Total	Absolute	0	0	0	1	7	112	930
		Relative	0.0%	0.0%	0.0%	0.1%	0.7%	10.7%	88.6%

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). The articles' citation counts and rankings were stored and analyzed. To verify correct execution of the Google Scholar parser, spot checks were performed.

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

			Number of Search Results						
			[0,1]	[2, 10]	[11, 50]	[51, 250]	[251, 1000]	[1001, 10000]	[10001, *]
Number of Query Terms	Single Words	Absolute	0	1	1	12	23	102	211
		Relative	0.0%	0.3%	0.3%	3.4%	6.6%	29.1%	60.3%
	Double Words	Absolute	166	89	54	27	11	3	0
		Relative	47.4%	25.4%	15.4%	7.7%	3.1%	0.9%	0.0%
	Triple Words	Absolute	345	5	0	0	0	0	0
		Relative	98.6%	1.4%	0.0%	0.0%	0.0%	0.0%	0.0%
	Total	Absolute	511	95	55	39	34	105	211
		Relative	48.7%	9.0%	5.2%	3.7%	3.2%	10.0%	20.1%

[3] All data was collected in November 2008

[4] The words for creating the search queries were extracted from an academic word list [16]

To identify interrelationships between citation counts and positions, the distribution of citation counts on the first and last positions were analyzed. This aimed to recognize whether articles with high/low citation counts occur more/less often in high/low positions. In addition, all results of the search queries were visualized to recognize patterns. This was performed for original citation counts and citation counts transformed to an ordinal scale. The transformation was performed for the results of each search query as follows: The lowest citation count was replaced with 0, the second lowest with 1 and so on (see Table 3).

The transformation was performed to ease the visualization process. Differences between graphs based on original and ordinal citation counts are illustrated in Figure 3 and Figure 4. By transforming citation counts the data's meaning changes slightly. In contrast to absolute citation counts, an ordinal citation count of, "5" means that his paper has the fifth lowest citation count of those articles in the result set. All graphs in this paper are based on ordinal data if not stated otherwise. Overall, a total of 3,122 graphs were created and inspected individually (1,561 graphs displaying original citation counts, and 1,561 graphs displaying ordinal citation counts).

Table 3: Transformation of Citation Counts

Original Data

	Result 1	Result 2	Result 3	Result 4	…
Query 1	593	18	5	5	
Query 2	485	6932	311	298	
…					

Transformed Data (Ordinal)

	Result 1	Result 2	Result 3	Result 4	…
Query 1	3	2	1	1	
Query 2	3	4	2	1	
…					

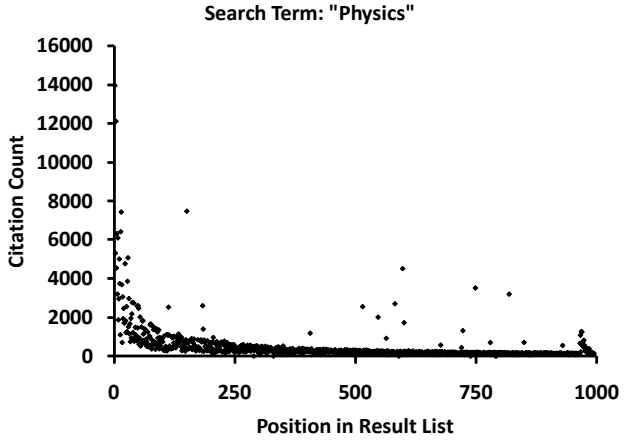

Figure 3: Visualization of Original Citation Counts

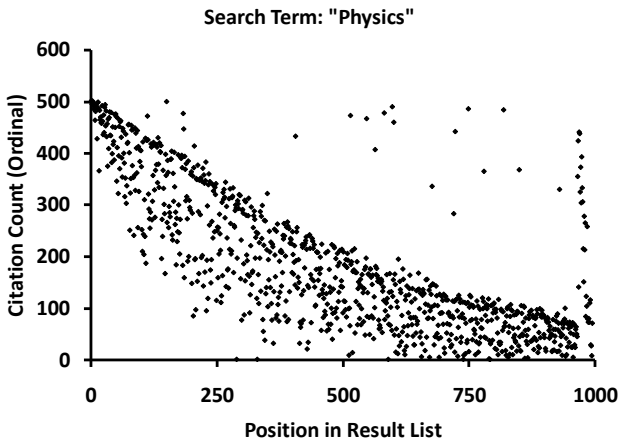

Figure 4: Visualization of Ordinal Citation Counts

5. RESULTS & INTERPRETATION

5.1 Citation Count Distribution

In Table 4 the distribution of articles' citation counts in comparison to their ranking in Google Scholar is listed for searches in the full text. 23.3% of all analyzed articles had zero citations, but only 14.7% of those articles ranked in position 1 to 10 had zero citations and only 10.8% of those articles in position 1. If citation counts do not impact the ranking, one would expect around 23% of zero-cited articles in a first position. In contrast, 16.7% of the articles ranked in position 1 had more than 1,000 citations, although these types of articles made up only 0.8% of the total articles. Overall, the data clearly shows that citation counts do have a significant impact on articles' ranking.

Table 4: Articles' Citation Count Distribution (Full Text)

	Citation Count							
	0	[1, 3]	[4, 10]	[11, 50]	[51, 150]	[151, 500]	[501, 1000]	[1001, *]
Position 1	10.8%	3.8%	7.6%	19.2%	15.6%	17.7%	8.6%	16.7%
Top 10	14.7%	8.0%	9.0%	18.6%	15.9%	17.2%	7.2%	9.4%
Total	23.3%	14.7%	14.9%	24.1%	13.0%	7.6%	1.6%	0.8%

The same analysis for title searches is presented in Table 5. Interestingly, 10.6% of the top 10 articles have zero citations although only 7.5% of all articles had zero citations. The reason for this became clear after examining the raw data. The search queries were created automatically by combining different words from a word list and some senseless search queries such as 'finish father' or 'excessive royalty' emerged. Therefore, 57.7% of the 1,050 search queries delivered only 10 or less results, and many of these results had few citations (see Table 2). If only articles in which search queries with more than 100 results are considered, the data would look similar to the data from Table 4. Additionally, 24.3% of those articles ranked in position 1 had more than 1,000 citations, although this type of article makes up only 0.7% of the total articles. Overall, the data confirms that citation counts have a definite and significant impact on an article's ranking. Differences between title and full text searches seem less significant.

Table 5: Articles' Citation Count Distribution (Title)

	Citation Count							
	0	[1, 3]	[4, 10]	[11, 50]	[51, 150]	[151, 500]	[501, 1000]	[1001, *]
Position 1	4.6%	9.6%	6.9%	11.5%	10.9%	19.3%	12.8%	24.3%
Top 10	10.6%	8.2%	6.3%	11.2%	15.4%	22.7%	10.9%	14.7%
Total	7.5%	15.9%	15.0%	28.0%	20.0%	11.2%	1.8%	0.7%

5.2 Graphs of the Means

Figure 5 illustrates the mean citation count per position for searches in the full text. It is clearly recognizable that a strong relationship exists between an article's citation count and its position. What stands out is the increase of the mean citation counts in the later positions. At first glance, it seems likely that outliers distort the citation counts of the last positions. In our dataset the mean citation count for position 1,000 was calculated from only 88 numbers[5]. By contrast, the mean citation count for position 1 was calculated from 1,050 numbers. Therefore, few but high outliers could have distorted the calculation. However, further analysis revealed that not only mean, but also median citation counts are significantly higher on the very last positions than in the positions before.

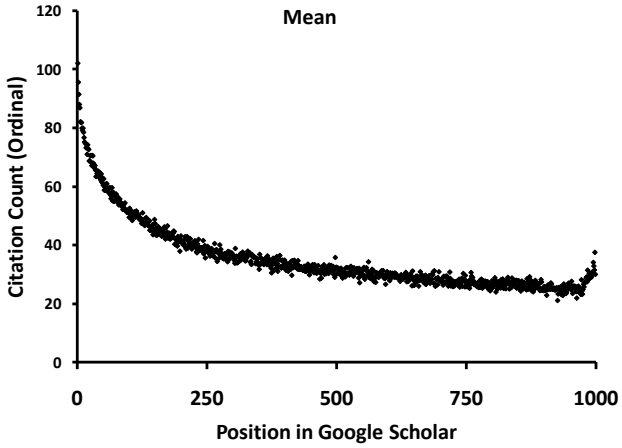

Figure 5: Mean Citation Counts (Full Text Search)

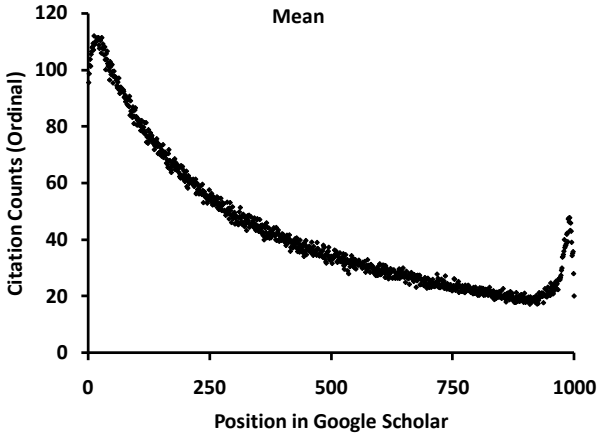

Figure 6: Mean Citation Counts (Title Search)

[5] There are two reasons for the low sample size of the last positions. First, not all search queries delivered 1000 results. Second, Google Scholar often displays slightly less than 1000 results even if there are 1000 or more results in its database.

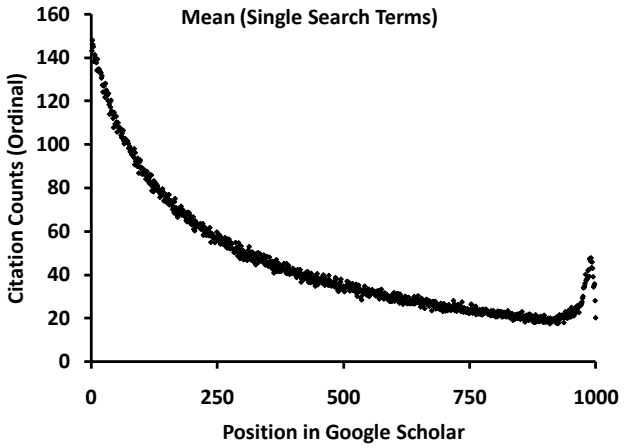

Figure 7: Mean Citation Counts (Title Search, Single Words)

The graph for title searches (see Figure 6) draws a similar picture. Only the lower mean citation counts in the first positions stand out. A closer examination explains this. For searches in the title with search queries consisting of two or three search terms, many results with few citation counts were returned by Google Scholar (see Table 2 and Table 5). This small sample size pushes the means in the first positions downwards. When search queries with less than 50 results are ignored, the search in the title presents a similar graph as the search in the full text. This is shown in Figure 7, which displays the means for the title search with single- word queries only.

Overall, all graphs show an almost perfect interrrelationship between an article's citation count and its position in Google Scholar. It could be assumed that other factors play either a minor role or none at all in Google Scholar's ranking algorithm. However, it stands out that the mean citation counts in the last positions significantely increase. We have no explanation for this phenomenon.

5.3 Graphs of Individual Search Queries

The graphs of the mean citation counts show a very clear interrelationship between an article's citation count and the way it is ranked by Google Scholar. However, in our previous study we discovered various patterns for individual search queries that differ from the graph of mean citation counts. Therefore, we analyzed the graphs of all individual search queries and discovered six different graph types.

5.3.1 Standard Graph

This type of graph (see Figure 8) looks as one would expect from what the research indicated so far: A strong interrelationship exists between a paper's citation count and its position in Google Scholar. As observed previously, the last positions often show a comparatively high citation count. Additionally, some significant outliers exist.

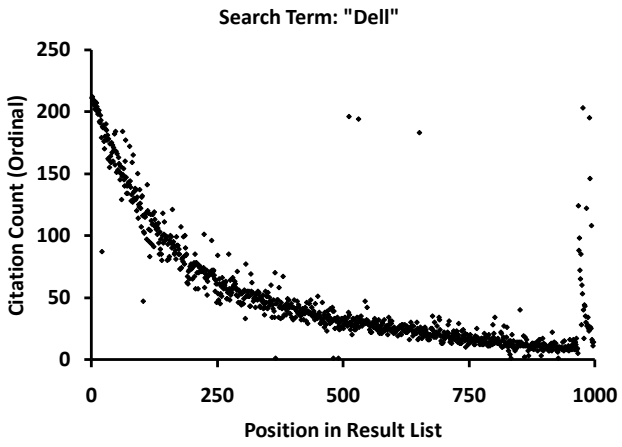

Figure 8: Standard Graph

5.3.2 Weak Standard Graph

This type of graph is similar to the standard graph, but correlation between citation counts and positions appears weaker (see Figure 9). The existence of this type of graph indicates that there are other important factors determining the position of an article in Google Scholar's results list.

5.3.3 No Pattern

This type of graph indicates no interrelationship of a paper's citation count and its position in Google Scholar at all (see Figure 10). It completely contradicts the overall observation that higher citation counts lead to a better ranking. We could not find any explanation of why this type of graph occurs. Apparently there are situations in which citation counts have no impact.

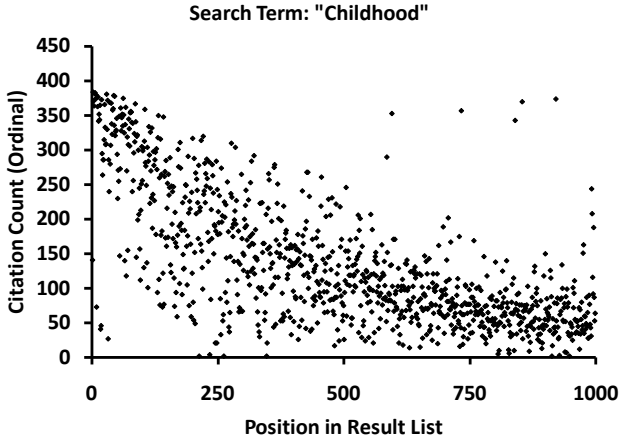

Figure 9: Weak Standard Graph

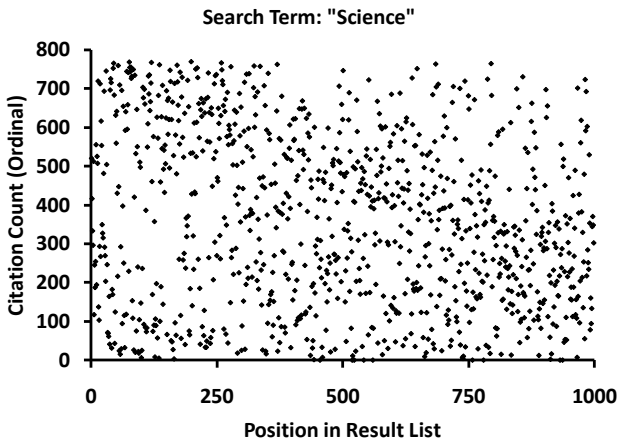

Figure 10: No Pattern Graph

5.3.4 Two in One Graph

This type of graph looks like a combination of two individual graphs (see Figure 11). At first glance it could be assumed that it occurs for search queries consisting of two words and Google Scholar calculates for each word, a separate result set and finally merges the two results sets. However, as shown in Figure 11, the 'two-in-one' graph also occurs for single word search queries.

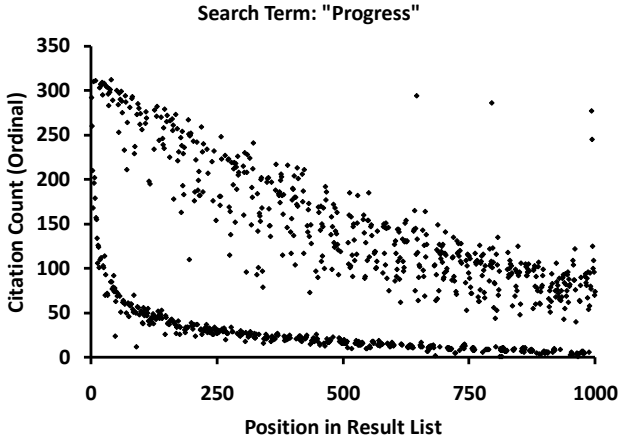

Figure 11: Two-in-One Graph

5.3.5 Interrupted Graph

This type of graph is intermittently interrupted (see Figure 12).

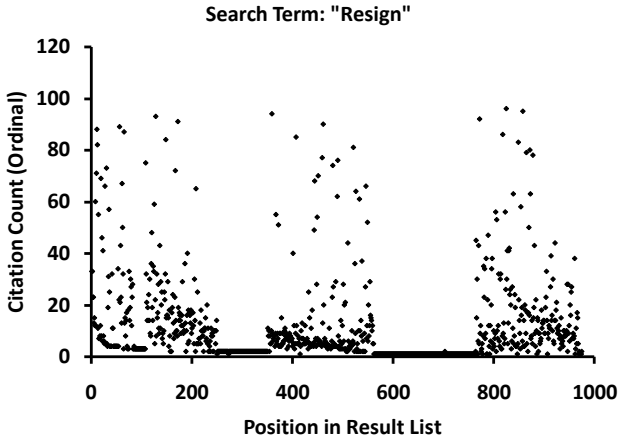

Figure 12: Interrupted Graph I

5.3.6 Combinations

Various graphs exist that look like a combination of the previously described graphs (see Figure 13).

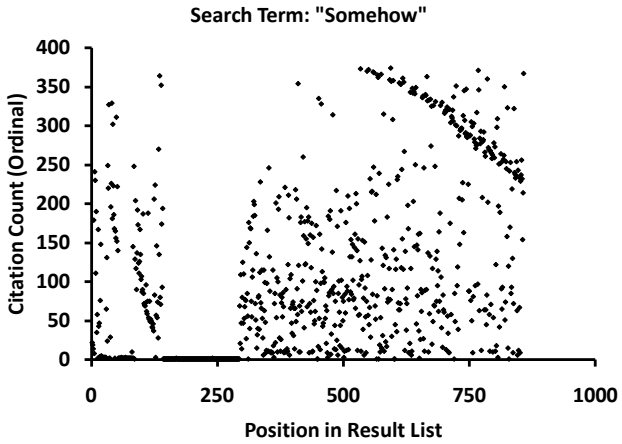

Figure 13: Combined Graph

5.3.7 Interpretation

While the graph of the means indicated an almost perfect interrelationship between citation counts and positions, the individual graphs do not show the same picture. Although citation counts do have a strong impact in most cases, it is not so in all cases.

Looking at the graphs resulting from full text searches, it stands out that only a minority of graphs equaled the standard graph. Most graphs were two-in-one or weak standard graphs or combinations. In addition, only few single-word search queries resulted in no pattern graphs, while most search queries consisting of two and three words resulted in no-pattern graphs. Since all search queries were created randomly we repeated the analysis with some 'realistic' search queries consisting of multiple words such as 'impact factor' or 'total quality management'. In these cases, the resulting graphs equaled more often the (weak) standard or two-in-one graph or combinations of them.

However, we want to emphasize that for any type of search query, any graph-type occurred at least once.

In contrast to full text searches, title searches resulted only in (weak) standard graphs and two-in-one graphs. Other graphs did not occur. In other words, Google Scholar's ranking algorithm for title searches *always* weighs citation counts heavily, while the algorithm for full text searches weighs citation counts heavily *most of the time*. This confirms our previous results that Google Scholar uses a (slightly) different ranking algorithm for title and full text searches. The reason is unknown.

6. SUMMARY AND OUTLOOK

Our study showed that an article's citation count does have a significant impact on the article's ranking in Google Scholar. The more citations an article has, the more likely it is displayed in a top position in Google Scholar's results list. This is true in most cases. While the results of title searches are always ranked heavily based on citation counts, Google sometimes makes exceptions in full text searches. This is especially the case for multi-word search queries, but not exclusively. We could not find

any reasons why Google seems to weight citation counts with different strengths. Here, further research is required as well as for the weight of other factors such as the age (upcoming paper see [14]).

In the final analysis, due to the strong weight on citation counts, Google Scholar is more suited when searching for popular standard literature than for searching gems, the latest trends, or papers whose authors are advancing views opposite to the mainstream. This is neither good nor bad, but users should be aware of it. Google Scholar also strengthens the Matthew Effect: articles with many citations will be more likely displayed in a top position, get more readers and receive more citations, which then consolidate their lead over articles which are cited less often. If Google Scholar should become only partly as popular for scientific articles as it is for web pages, there would be an even higher incentive for researchers to influence their article's citation counts; for instance via self citations or citation alliances.

7. REMARK

We would be delighted to share the Google Scholar parser software, including gathered data, with other researchers who wish to perform their own research or evaluate ours. Please send us an email if you are interested in the data or the software.

8. ACKNOWLEDGEMENTS

Our thanks go to Ammar Shaker for supporting the development of the Google Scholar parser. In addition we have to thank Dr. Wolfgang Lehmann for his advice. We regret that we had not the time to consider all of his feedback.

9. REFERENCES

[1] R. K. Merton, "The Matthew Effect in Science," *Science*, vol. 159, no. 3810, pp. 56–63, January 1968.

[2] (2006, August) Click Through Rate of Google Search Results - AOL-data.tgz - Want to Know How Many Clicks The no.1 Google Position Gets? Red Cardinal Blog. Red Cardinal Ltd.

[3] 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.

[4] J. Beel and B. Gipp, "Collaborative Document Evaluation: An Alternative Approach to Classic Peer Review," in *5th International Conference on Digital Libraries (ICDL'08)*, ser. Proceedings of World Academy of Science, Engineering and Technology, vol. 31, August 2008, pp. 410–413.

[5] 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.

[6] 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.

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

[8] 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.

[9] J. Bar-Ilan, "Which h-index? - A comparison of WoS, Scopus and Google Scholar," *Scientometrics*, vol. 74, no. 2, pp. 257–271, 2007.

[10] P. Jacso, "Google Scholar: the pros and the cons," *Online Information Review*, vol. 29, no. 2, pp. 208–214, 2005.

[11] 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.

[12] (2008) About Google Scholar. Website. Google Inc. [Online]. Available: http://scholar.google.com/intl/en/scholar/about.html

[13] J. Beel and B. Gipp, "Google Scholar's Ranking Algorithm: An Introductory Overview (Research in Progress)," to be published, 2009.

[14] J. Beel and B. Gipp, "Google Scholar's Ranking Algorithm: The Impact of Articles' Age (An Empirical Study)," in *Proceedings of 6th International Conference on Information Technology : New Generations (ITNG'09)*. IEEE, 2009.

[15] B. Gipp and J. Beel, "Scienstein: A Research Paper Recommender System," in *International Conference on Emerging Trends in Computing*. IEEE, 2009, pp. 309–315.

[16] S. Haywood. (2008) The Academic Word List. University of Nottingham. [Online]. Available: http://www.nottingham.ac.uk/ alzsh3/acvocab/wordlists.htm