

Google Scholar's Ranking Algorithm: An Introductory Overview

Jöran Beel

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

Bela Gipp

Otto-von-Guericke University
Department of Computer Science
ITI / VLBA-Lab / Scienstein.org
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. We performed the first steps to reverse-engineering Google Scholar's ranking algorithm and present the results in this research-in-progress paper. The results are: *Citation counts* is the highest weighed factor in Google Scholar's ranking algorithm. Therefore, highly cited articles are found significantly more often in higher positions than articles that have been cited less often. As a consequence, Google Scholar seems to be more suitable for finding standard literature than gems or articles by authors advancing a new or different view from the mainstream. However, interesting exceptions for some search queries occurred. Moreover, the occurrence of a search term in an article's title seems to have a strong impact on the article's ranking. The impact of search term frequencies in an article's full text is weak. That means it makes no difference in an article's ranking if the article contains the query terms only once or multiple times. It was further researched whether the name of an author or journal has an impact on the ranking and whether differences exist between the ranking algorithms of different search modes that Google Scholar offers. The answer in both of these cases was "yes". The results of our research may help authors to optimize their articles for Google Scholar and enable researchers to estimate the usefulness of Google Scholar with respect to their search intention and hence the need to use further academic search engines or databases.

Academic Search Engines, Google Scholar, Ranking Algorithm, Research in Progress

I. INTRODUCTION

With the increasing use of academic search engines, it becomes more important for academic authors to have their articles well ranked in these search engines in order to reach their audience. To optimize papers for academic search engines, such as Google Scholar or Scienstein.org, knowledge about the applied ranking algorithms is essential. For instance, if a search engine considers how often search terms occur in an article's full text, authors should use the most relevant keywords in their articles whenever possible to get their papers in a top position.

For users of academic search engines, knowledge about applied ranking algorithms is also essential in order to evaluate the robustness of the results. As pointed out, researchers do have an interest in having their articles displayed in top positions by search engines. Accordingly, users should know about the algorithms in order to

estimate the robustness and therefore the trustworthiness of the academic search engines' results.

Knowledge of ranking algorithms also enables researchers to estimate the usefulness of results in respect to their search intention. For instance, researchers intending to search for the latest trends in their field should use a search engine putting a 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 by authors advancing a view different from the majority, search engines putting high weight on citation counts might not give the best results.

This paper is structured as follows. First, an overview of related work is given including common ranking algorithms of academic search engines. Then, nine objectives are presented which were pursued by our research. This is followed by a section about the methodology and finally, results are presented. In the conclusion our interpretation of the results is presented and an outlook towards further research is given.

II. RELATED WORK

Due to different user needs, many academic databases and search engines let the user choose a ranking algorithm. For instance, ScienceDirect lets users choose between date and relevance¹, IEEE Xplore offers in addition, a ranking by title and ACM Digital Library lets users choose whether to sort results by relevance, date, alphabetically by title or journal, citation counts or downloads. However, these 'algorithms' can be considered of little worth since users can select only one ranking criteria and are not allowed to use a combination of them.

Google Scholar is one of the few academic search engines combining several approaches in a single algorithm². Several studies about Google Scholar exist. These studies include 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 indices such as the h-index [5] and the reliability of Google Scholar as a serious information source in general [6, 7].

However, 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 articles about Google Scholar's ranking algorithm. The only vague information about the

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

² Others are, for instance, *CiteSeer* and *Scienstein.org* [12, 13]

algorithm is from Google itself: 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.

III. RESEARCH OBJECTIVE

The research objective was to get a first impression on whether factors other than the four above-mentioned (full text, author, publication, and citations) are considered by Google Scholar and how much weight is put on each factor. The current research did not intend to give a final answer to all questions thoroughly, but to show which direction further research should go.

The following factors were researched:

- (1) Article's citation count
- (2) Article's age
- (3) Search term occurrence in an article's full text
- (4) Search term frequency in an article's full text
- (5) Search term occurrence in an article's title
- (6) Search term occurrence in author or publication name

In other words, we compared whether old/recent articles with high/low citation count and/or the search term occurring (frequently) in the title, full text and/or author/publication name occurring in the search query are more likely to be displayed in a top position by Google Scholar.

Since Google Scholar offers two basic search modes, namely a search in the full text and a search in the title, we also analyzed (7) whether the same ranking algorithm is used for both search modes. The next objective (8) was to analyze how rankings of articles retrieved via the ‘cited by’ and ‘related articles’ functions differ from those retrieved via normal keyword search. The research objective (9) was to analyze whether Google Scholar indexes text from figures and tables embedded as pictures in the articles.

IV. METHODOLOGY

A. Impact of Citation Counts

Google Scholar displays for each article its citation count in the result list. To obtain citation counts we developed a program to parse Google Scholar's website. 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 was performed twice, each time with 10 search queries. In the first run, the articles' full text was searched. In the second run it was the article's title only.

This resulted in 20 search queries, returning a total of 19,612 articles. The articles' citation counts and rankings were stored and analyzed. To verify correct execution of the Google Scholar parser, spot checks were performed.

To identify causal relationships and patterns between citation counts and rankings, all results were visualized. This was performed for the original citation counts and the citation counts transformed to an ordinal scale. The transformation was performed for the results of each search query as follows: The lowest citation counts were replaced with 1, the second lowest with 2 and so on (see Table 1). The transformation was performed to ease the visualization process. Differences between graphs with original and ordinal citation counts are illustrated in Figure 1 and Figure 2. All graphs in this paper are based on ordinal data if not stated otherwise.

TABLE I. 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	
...					

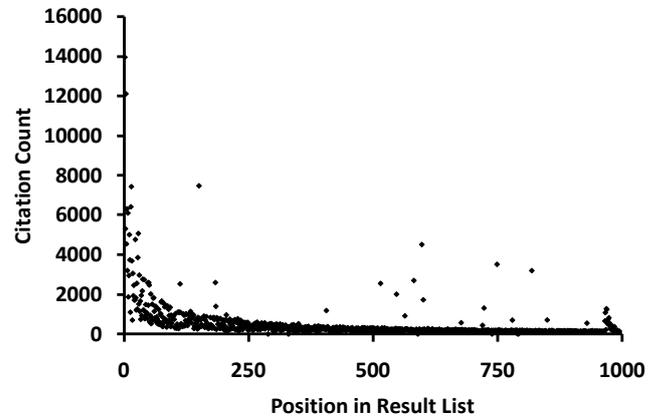


Figure 1. Visualization of Original Citation Counts (Search Term ‘Physics’)

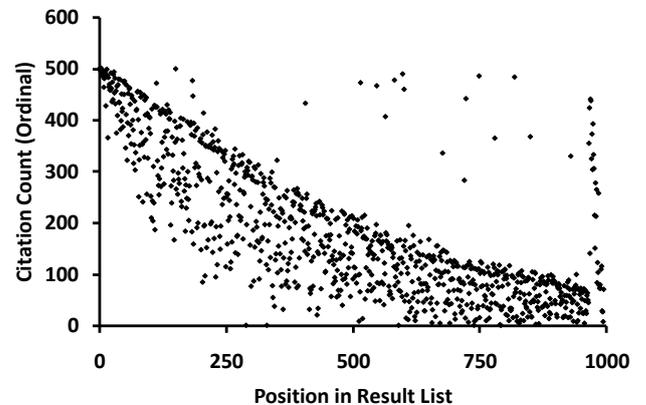


Figure 2. Visualization of Ordinal Citation Counts (Search Term ‘Physics’)

B. Impact of Age

To understand the impact that an article's age has on its ranking, 10 search queries were executed and the article's publication date for all results was retrieved by the Google Scholar parser. The search was performed in the full text and returned a total of 9,717 articles including publication year. The data was visualized individually (ten different graphs) and as a graph displaying the mean publication year of each position.

C. Impact of Search Term Occurrence in Full Text

We wanted to know whether a search term must occur in a document or if Google considers synonyms as well. To accomplish this, the results of 10 search queries were examined. For each search query, 12 articles from the result list were downloaded (4 from some of the first result pages, 4 from some of the middle pages and 4 from the last pages). It was examined whether the articles contained the search term at least once.

D. Impact of Search Term Frequency in Full Text

It would be plausible for Google Scholar to put high weight on the search term frequency in the full text (as Google does for web pages). A scientist searching for articles about 'RFID' probably would prefer a document containing the term 'RFID' fifty times more than a document containing it once. To determine whether the search term frequency in an article's full text impacts its ranking, the results of 5 search queries were analyzed. The full text for those results with the same citation counts were downloaded (when available) and the occurrences of the search terms counted. In addition, the relative search term frequency was calculated (search term count divided by total word count). The data was displayed and analyzed in a table and as a graph.

E. Impact of Search Term Occurrence in Title

To analyze whether the occurrence of a search term in an article's title has an impact on the article's ranking, results of 10 search queries were analyzed. It was compared how often the first ten results contained the search term in the title and how often the last ten results contained the search term in the title.

F. Difference between Search in Title and Search in Full Text

We assumed that result lists from searches in the title would equal the result lists from normal searches removed by the entries that do not have the search term in the title. In order to research whether Google is doing this, 10 search queries (executed as a search in the title and executed as a search in the full text) were analyzed. It was compared whether result lists of full text searches cleared by the entries not having the search term in the title, equaled the results of title searches.

G. Differences between 'Cited By', 'Related Articles' and Normal Keyword Search

To discover differences in the ranking algorithms of articles found via standard keyword search and 'cited by' and 'related articles' function, result lists of 10 search queries were analyzed. It was examined whether the result lists appeared similar or whether obvious differences occurred.

H. Indexing of Figures and Tables Embedded as Image

To examine if Google Scholar applies OCR to index text in images, 10 documents that included text as images were downloaded. Then, a search query was executed for text from the images that did not occur elsewhere in the text. It was then analyzed whether the document appeared in Google Scholar's result list.

I. Impact of Author and Journal Name

To analyze whether the existence of search terms in journal or author names have an impact, the first 20 results of 20 search queries were analyzed. The search queries consisted of words that were likely to be both part of a standard search query and part of an author or journal name. Words used among others: *brain*, *hammer*, *berry*, *black*, and *white*.

J. Remark

All data was collected in October 2008. We would like to explicitly point out that sample sizes were small and therefore the current research shall only be seen as a rough overview and introduction to Google Scholar's ranking algorithm to get an idea as to which direction further research should go. Other factors that might also have an impact on an article's ranking, such as authors' and publications' reputation were not researched, as the required information is not available via Google Scholar. A more exhaustive analysis follows in the upcoming papers [10] and [11].

V. RESULTS & INTERPRETATION

A. Impact of Citation Counts

The following graphs display for a particular search query the citation counts of a paper for each position of the results list. For instance Figure 3: Searching for 'Dell' in Google Scholar returns a results list in which the paper in position one has about 210 citations, while the paper in position 500 has about 40 citation counts.

All graphs show a clear interdependency between an article's citation count and the way it is ranked by Google Scholar. That shows us that the higher an article's citation count, the more likely it will be displayed in a top position. However, we discovered that three different types of graphs exist.

1) Standard Graph

This type of graph (see Figure 3) indicates a very strong dependency between a paper's citation count and its position in Google Scholar. One could assume that citation counts are practically the only significant factor for ranking research articles in Google Scholar's ranking algorithm.

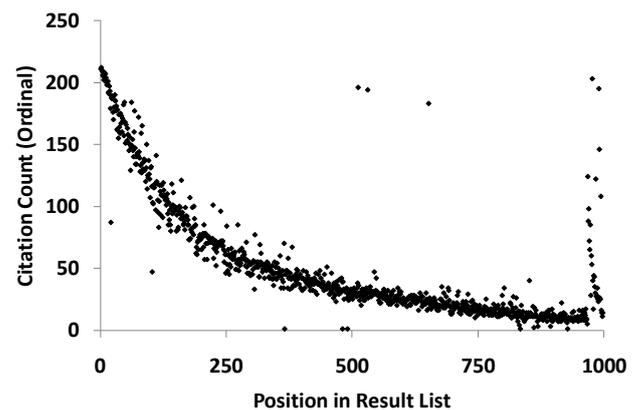


Figure 3. Standard Graph (Search Term 'Dell')

The last positions show comparatively high citation counts and other such significant outliers exist. Apparently, in these cases at least one other factor, unbeknown to us, had a significant impact on the ranking.

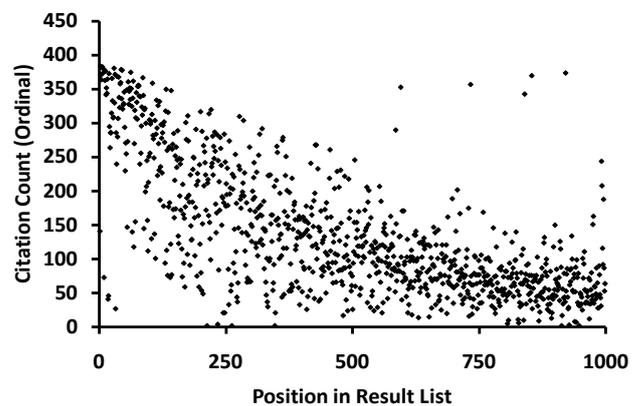


Figure 4. Weak Standard Graph (Search Term 'Childhood')

2) Weak Standard Graph

This type of graph is similar to the standard graph, but the dependency between citation counts and position appears weaker (see Figure 4). The existence of this type of graph indicates that there are

other important factors determining the position of an article in Google Scholar's result list.

3) Two in One Graph

This type of graph looks like a combination of two individual graphs (see Figure 5). It could mean that Google Scholar somehow retrieves two different result lists, ranks each by citation counts and then simply merges the two lists.

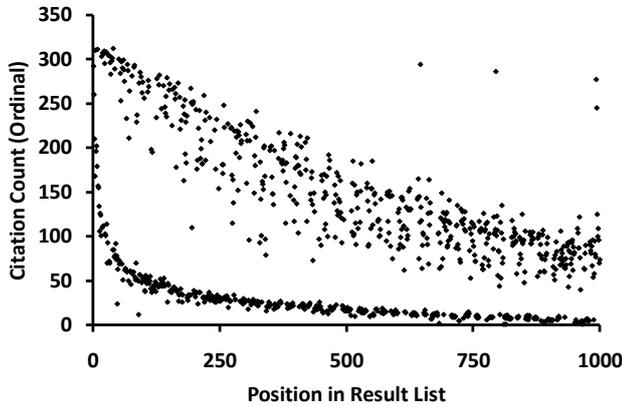


Figure 5. Two-in-One Graph (Search Term 'Progress')

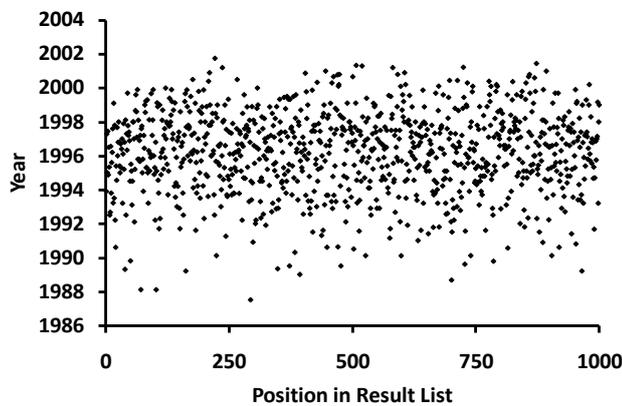


Figure 6. Mean Publication Year per Position

B. Impact of Age

Since citation counts apparently do have a strong impact on Google Scholar's rankings, one could assume that older articles are found more often in first position, since older articles naturally have had more time to be cited. However, our research indicates that articles in the top positions are not necessarily older than articles in the later positions (see Figure 6). The graph shows the mean age of an article in relation to its ranking. It appears that Google Scholar weighs recent articles stronger than older articles in order to compensate for the Matthew effect.

C. Impact of Search Term Occurrence in Full Text

In all analyzed full texts, the search terms that were used occurred at least once in the text. Accordingly, it can be assumed that Google Scholar abides strictly to an article's text and does not consider synonyms.

D. Impact of Search Term Frequency in Full Text

The results of our analysis were not as expected: we found no direct relationship between articles' relative or absolute search term

frequencies in the full text and their ranking in Google Scholar (see Table 2, Figure 7 and Figure 8). This means that an article containing a search term multiple times is not more likely to be displayed in a top position than an article containing the search term only once. Reasons for this can only be speculative. It could be that Google Scholar wants to treat all articles equally. Because Google Scholar does not have access to all full texts of the articles listed in its database, it could be sensible to put little or no weight on the search term frequency in the full text.

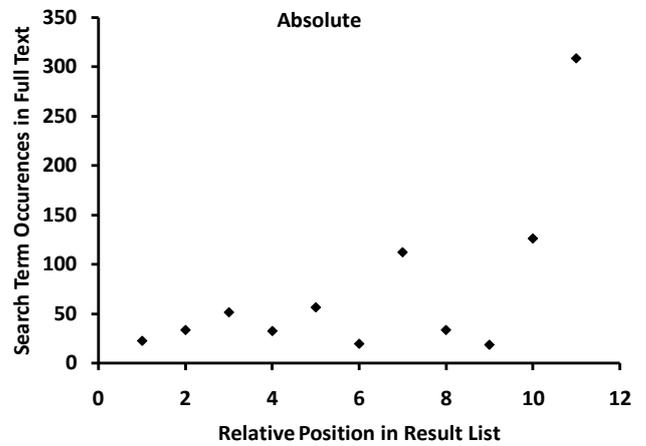


Figure 7. Absolute Search Term Count in Full Text

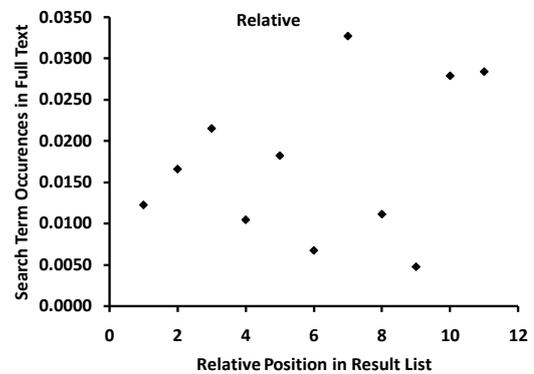


Figure 8. Relative Search Term Count in Full Text

TABLE II. SEARCH TERM COUNT IN FULL TEXT OF ARTICLES WITH 4 CITATIONS (SEARCH TERM: 'RFID')

Rel. Position	Year	Cita-tions	Keyw. Count	Word count	Rel. Keyw. Count
1	2003	4	22	1794	0.0123
2	2005	4	33	1987	0.0166
3	2006	4	51	2368	0.0215
4	2006	4	32	3060	0.0105
5	2005	4	56	3071	0.0182
6	2004	4	19	2817	0.0067
7	2005	4	112	3422	0.0327
8	2004	4	33	2962	0.0111
9	2004	4	18	3776	0.0048
10	2005	4	126	4514	0.0279
11	2006	4	309	10875	0.0284

However, it could be that our results cannot be generalized due to the small sample size. Two explanations exist why our results and interpretations might be incorrect. In all researched articles the search term occurred at least 12 times. It could be that Google Scholar treats documents equally as soon as a certain keyword frequency is

exceeded. Probably even more important is the fact that all researched articles had the search term in the title. It could be that Google Scholar ignores search terms in the full text if they already occur in the title.

E. Impact of Search Term Occurrence in Title

The results thus far give us reason to assume that the existence of a search term in an article's title has a definite impact on the article's ranking. Further research confirmed this. 86% of the articles listed in the top 10 of result lists contained the search term in the title, while only 26% of the last 10 positioned articles contained the search term in their title. This indicates that the words in the title have a high impact on an article's ranking.

F. Impact of Author and Journal Name

During our research we realized that if authors exist whose names are identical to the search term or parts of it, their articles are likely to be displayed in a top position in the results list. For instance, a search for 'white LED' returns in first position an article about 'The Federalists: a study in administrative history' which has nothing to do with white LEDs but was written by LED White³. This is true for journal and conference names as well. In the top 100 of a search for 'arteriosclerosis and thrombosis cure' 74 articles occur about various (medical) topics from the Journal 'Arteriosclerosis, Thrombosis, and Vascular Biology'.

G. Difference between Search in Title and Search in Full Text

Our analysis indicates that Google Scholar is using slightly different algorithms for searches in the title and searches in the full text. Although title searches return results lists similar to full text searches (eliminated by the entries not including the search term in the title), they were not exactly as expected. Figure 9 illustrates this with the example of the search for 'impact factor'.

In this example, the results list of the title search does not include the results 2, 3, 7 and 10 from the full text search. This is plausible because those articles did not include the term 'impact factor' in their title. However, results 12 and 11 are in opposite order and result 4 is placed completely differently than expected. This was similar for all search queries, which means that Google Scholar uses slightly different algorithms. We can see no obvious reason for this.

Search in Full Text	Search in Title
Result 1	Result 1
Result 2	Result 2
Result 3	Result 3
Result 4	Result 5
Result 5	Result 6
Result 6	Result 7
Result 7	Result 8
Result 8	Result 9
Result 9	Result 10
Result 10	Result 12
Result 11	Result 11
Result 12	Result 13
Result 13	Result 4
Result 14	Result 14
...	...

Figure 9. Full Text Search vs. Title Search

H. Differences between 'Cited By', 'Related Articles' and Normal Keyword Search

Apparently, Google Scholar uses a different algorithm for sorting the 'Cited By' results list than for the results of the standard search. It seems as if this algorithm puts a very high weighting on citation counts. For the 'related articles' function, Google Scholar again uses

³ The author's real name is *Leonard Dupee White* but Google Scholar failed recognizing his name properly. However, this example illustrates that Google Scholar puts a high weighting on author names.

another algorithm. At first glance, the ranking algorithm does seem to put low weighting on citation counts.

I. Indexing of Figures and Tables Embedded as Image

Text which is embedded into a document via images and has to be recognized via OCR is not indexed by Google Scholar. This statement is true for documents containing mostly text and only some images. We did not analyze whether Google Scholar indexes complete scans of documents via OCR, which would probably lead to indexing the whole text of a document.

VI. CONCLUSION AND OUTLOOK

An article's citation count does have significant impact on its ranking in Google Scholar. That means that Google Scholar is more suitable when searching for standard literature and less suitable when searching for gems 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 to be displayed in a top position, get more readers and receive more citations, which then consolidate their lead over articles that 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.

A good title is important for a scientific paper anyway, but the more Google Scholar is used, the more an author should think about an article's title in detail. Beside citation counts, the title seems to be one of the most important factors for the ranking algorithm. Therefore, it is just as important that a title makes the reader curious, and includes the most relevant keywords to achieve high positions in Google Scholar.

Since Google Scholar does not consider synonyms, users should think carefully about the terms they search for. Otherwise they could miss out on relevant documents. In addition, authors should think carefully about the terminology they use in their articles because this might decide whether their articles are found by Google Scholar users or not. If our initial research is correct and the frequency of keyword use has very little or no impact, it might be advisable for authors to use a variety of different terms and synonyms in their articles. As a consequence, their articles might be less readable, but their chances to be found would increase.

Authors embedding figures and tables in their documents as images should reconsider. Google Scholar seems not to use OCR to recognize text in images. Relevant keywords may not be indexed from Google Scholar if they are displayed as an image. Authors should use vector graphics and tables with 'real' text instead.

Overall, this study has raised more questions than it has answered. We showed that differences exist between the algorithms for a keyword search in the full text, in the title, the 'related articles' and 'cited by' functions. However, it is not clear *why* Google Scholar uses different algorithms or *how* these algorithms differ exactly. Likewise, it remains unclear how strong the impact of citation counts actually is and why the graphs show different patterns. Most importantly, the exact weighting of the different factors remains unclear. Further research for all research objectives is required.

VII. SUMMARY

Academic search engines play an important role in searching for scientific articles. To maximize their effectiveness, users and authors should be aware of how search results are ranked. Most academic

databases such as the IEEE Xplore offer multiple but simple ranking algorithms and let the users choose only one in which they can apply.

Google Scholar, one of the major academic search engines, combines several factors. The exact algorithm is unknown. As a consequence, users do not know to what extent Google Scholar can satisfy their search intention and authors do not know how to prepare their papers for a good ranking.

We performed the first steps to reverse-engineering Google Scholar's ranking algorithm with the following results:

1. Overall, Google Scholar's ranking algorithm relies heavily on an article's citation count. As a result, Google Scholar strengthens the Matthew effect and is more suitable when searching for standard literature than gems, the latest trends, or articles by authors advancing a different view from the mainstream. Should Google Scholar become as popular for academic articles as it is for websites, the ranking algorithm will create further incentives for scholars to actively influence, or manipulate their citation counts.
2. Google Scholar's ranking algorithm puts a high weighting on words in the title. Knowing this, authors should think carefully about the title they give their articles. All relevant keywords should be included and it might be sensible to choose a long title.
3. Google Scholar considers only words that are included in an article, no synonyms. For this reason, users of Google Scholar should perform searches not only for one keyword but also for its synonyms. Otherwise they will miss out on relevant documents.
4. It appears as though Google Scholar does not put a weighting on the frequency in which search terms occur in the full text. This means that an article will not be ranked higher for a certain search just because the search term occurs more often in the full text. Because of this, it could be beneficial for authors to abstain from a strict terminology in their articles and use more synonyms. This would make their documents less readable but more retrievable in Google Scholar.
5. Google Scholar seems to weigh recent articles stronger than older articles which in turn, could compensate for the Matthew Effect.
6. Google Scholar is not indexing text embedded via images. Authors should avoid inserting tables, diagrams and figures as images (.png, .gif, .jpg, etc.) but use vector graphics and real text instead.
7. Google Scholar uses different ranking algorithms for a keyword search in the full text, keyword search title, the 'related articles' function and the 'cited by' function.
8. Google Scholar's ranking algorithm puts a high weighting on author and journal names. Users should be aware of this because it could distort the result list. In addition, it could also be beneficial for an author to publish in a journal whose

name includes keywords relevant to the article's content. The impact of an authors' and journals' reputation on the ranking has not been researched yet.

The research undertaken can only be seen as a first step. Many questions have remained unanswered. Further research is required, with more sample data and more analysis in order to get a comprehensive picture of Google Scholar's ranking algorithm.

VIII. ACKNOWLEDGEMENTS

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

REFERENCES

- [1] Yang, Kiduk and Meho, Lokman I., "Citation Analysis: A Comparison of Google Scholar, Scopus, and Web of Science," *Proceedings 69th Annual Meeting of the American Society for Information Science and Technology*, 2006.
- [2] J. Bailey et al., "Search Engine Overlaps: Do they agree or disagree?," *Second International Workshop on Realising Evidence-Based Software Engineering*, 2007, p. 2.
- [3] William H. Walters, "Google Scholar coverage of a multidisciplinary field," *Information processing & management*, vol. 43, 2007, 1121-1132.
- [4] John J. Meier and Thomas W. Conkling, "Google Scholar's Coverage of the Engineering Literature: An Empirical Study," *The Journal of Academic Librarianship*, vol. 34, 2008.
- [5] Judit Bar-Ilan, "Which h-index? — A comparison of WoS, Scopus and Google Scholar," *Scientometrics*, vol. 74, 2007, 257-271.
- [6] Péter Jacsó, "Google Scholar: the pros and the cons," *Online Information Review*, vol. 29, 2005, 208-214.
- [7] White, Bruce, "Examining the claims of Google Scholar as a serious information source," *New Zealand Library & Information Management Journal*, vol. 50, 11-24.
- [8] Google Scholar, "About Google Scholar"; <http://scholar.google.com/intl/en/scholar/about.html>.
- [9] Beel, Jöran and Gipp, Bela "Collaborative Document Evaluation: An Alternative Approach to Classic Peer Review," *Proceedings of World Academy of Science, Engineering and Technology*, Vienna: 2008, vol. 31, 410-413. Available at: www.sciencetechnology.org
- [10] Beel, Jöran & Gipp, Bela, "Google Scholar's Ranking Algorithm: The Impact of Citation Counts (An Empirical Study)", in Proceedings of the 3rd International Conference on Research Challenges in Information Science. 2009. *IEEE*. Available at: www.sciencetechnology.org
- [11] Beel, Jöran & Gipp, Bela, "Google Scholar's Ranking Algorithm: The Impact of Articles' Age (An Empirical Study)", in Proceedings of 6th International Conference on Information Technology: New Generations, 2009. *IEEE*. Available at: www.sciencetechnology.org
- [12] B. Gipp and J. Beel, "Scienstein: A Research Paper Recommender System", in *Proceedings of International Conference on Emerging Trends in Computing (ICETiC'09)*, 309-315. 2009. Available at: www.sciencetechnology.org
- [13] J. Beel and B. Gipp, "The Potential of Collaborative Document Evaluation for Science," *Springer's Lecture Notes in Computer Science (LNCS)*, vol. 5362, International Conference on Asian Digital Libraries (ICADL) Springer, 2008. Available at: www.sciencetechnology.org