2017

A Proposed Improvement to Google Scholar Algorithms Through Broad Topic Search Emergent Research Forum Paper

Matthew Russell Kearl
Dakota State University

Cherie Noteboom
Dakota State University

Deb Tech
Dakota State University

A Proposed Improvement to Google Scholar Algorithms Through Broad Topic Search Emergent Research Forum Paper

Conference Paper · August 2017

CITATIONS
0

READS
49

3 authors, including:

Cherie Bakker Noteboom
Dakota State University
33 PUBLICATIONS **52** CITATIONS

SEE PROFILE

A Proposed Improvement to Google Scholar Algorithms Through Broad Topic Search

Emergent Research Forum Paper

Matthew Kearl
Dakota State University
Kearl@dixie.edu

Dr. Cherie Bakker Noteboom
Dakota State University
Cherie.Noteboom@dsu.edu

Dr. Deb Tech
Dakota State University
Deb.Tech@dsu.edu

Abstract

Google Scholar uses ranking algorithms to find the most relevant academic research possible. However, its algorithms use an exact keyword match that excludes synonymous search terms that may be overlooked or neglected by researchers. This paper aims to improve on the current Google Scholar Search System by allowing a broad topic search algorithm to diversify and allow synonymous search terms to be included and ranked with other results. The authors propose a Design Science method to improve the Google Scholar Search System by developing a broad topic prototype that will add synonymous keywords into Google Scholar ranking algorithms. The results from twenty users will be evaluated by means of Mean Reciprocal Rank and Discounted Cumulative Gain. This improvement will introduce a modern approach to academic search engines systems, and to allow researchers who overlook potential search queries, an improved core topic diversity, quality, and discoverability of published research.

Keywords

Google Scholar, Ranking Algorithms, Broad Topic Search, Academic Search Engine

Introduction

In 2009, it was estimated that the number of published scientific research papers (SRP) exceeded 50 million (Jinha 2010). In 2015, it was estimated that the global scientific output doubled every nine years (Bornmann and Mutz 2015). Today, the need for effective search tools has never been greater. During the last decade, Google Scholar (GS) has been gaining traction as a critical tool for research and discovery of new SRP. GS uses algorithms to rank and return relevant results from a user's search query. Depending on these algorithms, relevant or less relevant results will be returned to the user. Although Google Scholar does offer alternative query links after the Search Engine Results Page (SERP) (Shetty 2016), the system does not include those within the SERP, making it difficult to find the best results from all potentially related terms or synonymous queries. Furthermore, when compared to modern web search engines today, GS algorithms have fallen behind in broad topic match algorithms, and are primarily based on citation count and exact keyword match, and often do not provide as relevant results as it might otherwise if algorithms were updated (Amolochitis 2014; Beel and Gipp 2009b; Hasson et al. 2014).

The problem with GS ranking is that citation count and exact keyword match signals are too strong and result in poor initial results to users. Because of this, ranking signals must be reexamined and modern algorithms built to better sort the SERP. This research aims to investigate current algorithm advantages and pitfalls, then proposes a new ranking method to be developed and tested for relevance against current ranking models. It is theorized that a new ranking method based on broad topic search in combination with current citation count algorithms, will produce more relevant search results to users.

Related Research

Like modern web search engines, GS uses web crawlers to discover new scientific content online that users may conduct search queries and explore results. It does this through a process that saves page content to databases called indexes. Specific evaluated areas of indexes are known as ranking signals. These specific ranking signals are fed into ranking algorithms that sort indexes into the most relevant results from high to low (De Winter et al. 2014). These indexed ranking signals often are evaluated at different levels of importance by ranking algorithms. Some of these include: keywords found in abstracts, body text, titles, figures, publication names, author names, author keywords, file names, subheadings, annotations, and metadata (Marks and Le 2016). Other signals examined include citation count, date or age of publication, author or publication reputation, and calculated h-index (Beel et al. 2009). These signals give ranking algorithms a clearer view on what the SRP is about and if it is of high quality.

In GS, the primary signal used to indicate quality is inbound citation count (Beel and Gipp 2009a). This is calculated by examining other works found in the database that cite the original SRP as a source. This democratic process of voting for others through citations (Page et al. 1999) acts as a signal for high-quality content (Martin-Martin et al. 2017) and encourages ranking algorithms to rank high on the SERP (Ale Ebrahim et al. 2014). However, many critics have argued that this method strengthens the Matthew Effect (Al-Hattab 2016) of the highly cited SRP receiving more citations and visibility than new emerging research (Martín-Martín et al. 2016). This is further strengthened by algorithm changes that encourage older SRP with high citation to rank higher (Verstak et al. 2014).

Furthermore, it has been shown that citation count can also be manipulated (Delgado López-Cózar et al. 2014) or spammed (Beel and Gipp 2010) through fictitious citation references linked between numerous indexed fabricated SRP's (Labbé 2010). Because of the Matthew effect and ease of manipulating references, citation count should not be the primary focus of ranking algorithms.

The second signal used in GS to find relevant content is exact keyword search (Beel and Gipp 2009b). When keywords are typed into a search query, indexes are searched for content matching the search term. Ranking algorithms will give higher or lower scores to each based on signal location in the index. For example, a search term found in an SRP title will rank higher than a single search term found in a figure description of another SRP. Location of keywords dictates the strength of the signal. The problem with exact match is that it does not allow synonymous keywords to be found even in high signal areas such as a title. For example, a simple experiment shows that a search query for article titles containing the words "female pay gap" will result in a completely different SERP than, "women's pay gap". Both queries should result in similar SRP, but if the researcher does not consider all synonymous key phrases, high-quality research could be missed. This means that to be found, authors must be certain to use every variation of keywords or potentially miss being discovered in GS. Even simple variations such as plural (police officers) and singular (police officer) key terms result in different SERP.

Because of this, a new field of academic search engine optimization has emerged where authors take advantage of these flaws and consciously optimize content so that it has a higher likelihood of ranking high in GS (Beel et al. 2009). Although it is a good practice to make research accessible and more visible (Ebrahim 2015; Kenny 2011), over-optimized results through keyword stuffing may prevent important works from being found by pushing them lower in the SERP.

Various approaches have been suggested to remedy some of these issues. First, the signal of paper age or a time depreciation score would allow old SRP with high citations to lose strength over time and allow higher ranking of new SRP (Amolochitis et al. 2013). Second, a publication venue signal, would rank higher quality conferences and journals above those of low quality (Hasson et al. 2014). Third, a term frequency heuristic signal would look at frequency, placement and relative distance of key terms (Amolochitis 2014). Although these novel solutions propose excellent methods of increasing quality of the SERP, none address the problem of limited results due to exact keyword match, and the exclusion of potential long tail keyword searches (Dennis 2016).

As shown below in figure 1, long tail search terms refer to the numerous niche keywords that a user might not consider when performing a search. These are synonymous keywords that might be used instead of another, combined with others, or alluded to in meaning. Because long tail search terms might not be entered as frequently, those results will receive low visibility on the SERP no matter the quality.

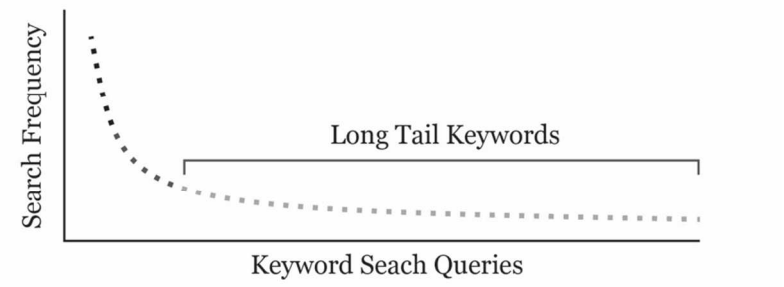

Figure 1. Long tail keywords with low SERP visibility (Skiera et al. 2010)

Modern web search engines have solved high-quality content discoverability issues through broad topic search or what is known as semantic topic clustering (Kong et al. 2016). This is where a search query is associated with a topic, and the topic includes all results for synonymous keywords and phrases.

The remainder of this paper aims to explore how integrating a broad topic search into GS results will enrich the SERP and increase high-quality results.

Methodology

The intention of this paper is to follow the design science research methodology (Peffers et al. 2007; Von Alan et al. 2004) of developing a new prototype application to demonstrate the use of broad topic search in Google Scholar. This is accomplished through design science methodology: (a) designing a proposal and making an awareness to the problems that exist in GS, (b) suggesting a solution and proposing a tentative broad topic analysis design, (c) future development and implementation of the design, (d) developing an evaluation and feedback plan, and (e) sharing results and conclusions (Vaishnavi and Kuechler 2004).

To explore the effects of broad topic-based search on GS, a proposed system would need to be developed to analyze search queries and include synonymous terms to be used to search indexes. As seen in figure 2, five processes are suggested that will allow the transformation of search terms into a topic-based SERP, which would be used as a foundation for a developed prototype.

Process P1 is used to capture the search queries entered in form fields from the user interface.

Process P2 will prepare the data for broad topic analysis. In this phase, nouns and adjectives are identified and saved to later find synonymous meaning, while filler keywords such as “and”, “the”, “a”, and “are” are ignored.

In process P3, saved keywords will be used to identify synonymous meaning. Modern search engines already use specialized algorithms to determine synonymous words. Therefore, for simplicity sake and to obtain the most robust data available, the Google AdWords Keyword Planner and the Onelook.com API will be used to identify similar-meaning long tail search queries.

In process P4, the ranking algorithm will use each of the newly identified queries to produce a combined list of GS sorted results. For example, if P3 determines there are eight synonymous search terms, then results from all searches will be combined by selecting the first result from all eight sets, then aggregating second results to the list and so forth. This should compile a diversified SERP with existing GS ranking algorithms. The primary purpose of this process is to diversify or expand results based on long tail keywords that researchers may neglect.

Figure 2. Broad Topic Analysis Processes

If this proposed system is to be effective, in process P5, an expanded SERP will be presented to the user including results from synonymous search terms generated by the new system. To measure the value and quality of ranked SERP, a group of academic researchers will be given a niche topic in their fields, and asked to organically select a search phrase of their choice, which the system will return twenty results or two pages of results. The authors felt that twenty SRP were sufficient since only 91% of users do not click beyond the first page of 10 results (Van Deursen and Van Dijk 2009). Each individual will then click on various SRP of their choice from the SERP, that will later be evaluated through Mean Reciprocal Rank (Craswell 2009) and Discounted Cumulative Gain (Dupret 2011). These will be used to assess the success of each set of results. Additionally, a qualitative survey will be used to determine the quality of the results and explore additional potential recommendations for improvement. To enhance the rigor and validity of the system, a group of 30 candidates will be used to evaluate 10 niches each, all with their own search phrases. By comparing all 10 niches and results across the 30 candidates, this designed system hopes to demonstrate that broad topic search is not only an effective feature in modern web search engines, but can also be applied to academic search engines.

Future Research

GS ranking algorithms are relatively undeveloped when compared to modern algorithms. To limit the Matthew Effect in process P4, additional algorithms such as time depreciation, venue location, and term frequency signals might be used to enhance the SERP quality. Furthermore, research into an author rank signal based partially on h-index (Bar-Ilan 2008), could gauge the quality of an author. Lastly, this paper focuses primarily on GS algorithms, however its methods could be further implemented and evaluated on other similar academic search engines.

Conclusion

This design science approach in adapting broad topic analysis to GS results overcomes its problem of exact keyword match and the need for multiple search queries with long tail keywords. It allows for citation count algorithms to be applied to a wide range of topic-based search phrases rather than one solitary query. This increases discovery of research in all areas of the topic and will result in more high-quality research results in academic search engines.

References

Al-Hattab, F.M.F. 2016. "An Efficient Ranking Algorithm for Scientific Research Papers." Zarqa University-Jordan.

Ale Ebrahim, N., Salehi, H., Embi, M.A., Habibi, F., Gholizadeh, H., and Motahar, S.M. 2014. "Visibility and Citation Impact," *International Education Studies* (7:4), March 30 2014.

Amolochitis, E. 2014. "Algorithms for Academic Search and Recommendation Systems," in: *Electronic Systems*. Aalborg University: Videnbasen for Aalborg UniversitetVBN, Aalborg UniversitetAalborg University, Det Teknisk-Naturvidenskabelige FakultetThe Faculty of Engineering and Science.

Amolochitis, E., Christou, I.T., Tan, Z.-H., and Prasad, R. 2013. "A Heuristic Hierarchical Scheme for Academic Search and Retrieval," *Information Processing & Management* (49:6), July 31 2013, pp. 1326-1343.

Bar-Ilan, J. 2008. "Which H-Index?—a Comparison of Wos, Scopus and Google Scholar," *Scientometrics* (74:2), 2008, pp. 257-271.

Beel, J., and Gipp, B. 2009a. "Google Scholar's Ranking Algorithm: The Impact of Citation Counts (an Empirical Study)," *Research Challenges in Information Science, 2009. RCIS 2009. Third International Conference on*: IEEE, pp. 439-446.

Beel, J., and Gipp, B. 2009b. "Google Scholar's Ranking Algorithm: An Introductory Overview," *Proceedings of the 12th International Conference on Scientometrics and Informetrics (ISSI'09)*: Rio de Janeiro (Brazil), pp. 230-241.

Beel, J., and Gipp, B. 2010. "Academic Search Engine Spam and Google Scholar's Resilience against It," *Journal of electronic publishing* (13:3).

Beel, J., Gipp, B., and Wilde, E. 2009. "Academic Search Engine Optimization (Aseo) Optimizing Scholarly Literature for Google Scholar & Co," *Journal of scholarly publishing* (41:2), pp. 176-190.

Bornmann, L., and Mutz, R. 2015. "Growth Rates of Modern Science: A Bibliometric Analysis Based on the Number of Publications and Cited References," *Journal of the Association for Information Science and Technology* (66:11), pp. 2215-2222.

Craswell, N. 2009. "Mean Reciprocal Rank," in *Encyclopedia of Database Systems*. Springer, pp. 1703-1703.

De Winter, J.C., Zadpoor, A.A., and Dodou, D. 2014. "The Expansion of Google Scholar Versus Web of Science: A Longitudinal Study," *Scientometrics* (98:2), pp. 1547-1565.

Delgado López-Cózar, E., Robinson-García, N., and Torres-Salinas, D. 2014. "The Google Scholar Experiment: How to Index False Papers and Manipulate Bibliometric Indicators," *Journal of the Association for Information Science and Technology* (65:3), pp. 446-454.

Dennis, J. 2016. "Search Engine Optimization and the Long Tail of Web Search," in: *Department of Linguistics and Philology*. Uppsala University.

Dupret, G. 2011. "Discounted Cumulative Gain and User Decision Models," *International Symposium on String Processing and Information Retrieval*: Springer, pp. 2-13.

Ebrahim, N.A. 2015. "Optimize Your Article for Search Engine." Retrieved Feb 1, 2017, from https://works.bepress.com/aleebrahim/110/

Hasson, M.A., Lu, S.F., and Hassoon, B.A. 2014. "Scientific Research Paper Ranking Algorithm Ptra: A Tradeoff between Time and Citation Network," *Applied Mechanics and Materials*: Trans Tech Publ, pp. 603-611.

Jinha, A.E. 2010. "Article 50 Million: An Estimate of the Number of Scholarly Articles in Existence," *Learned Publishing* (23:3), pp. 258-263.

Kenny, L. 2011. "Get Your Work Noticed: How Authors Can Help Readers to Find Annals Papers Online," in: *Oxford University Press*. BOHS.

Kong, J., Scott, A., and Goerg, G.M. 2016. "Improving Semantic Topic Clustering for Search Queries with Word Co-Occurrence and Bigraph Co-Clustering." Google Inc.

Labbé, C. 2010. "Ike Antkare One of the Great Stars in the Scientific Firmament," *International Society for Scientometrics and Informetrics Newsletter* (6:2), July 3 2012, pp. 48-52.

Marks, T., and Le, A. 2016. "Increasing Article Findability Online: The Four C's of Search Engine Optimization,").

Martín-Martín, A., Orduna-Malea, E., Ayllón, J.M., and López-Cózar, E.D. 2016. "Back to the Past: On the Shoulders of an Academic Search Engine Giant," *Scientometrics* (107:3), pp. 1477-1487.

Martin-Martin, A., Orduna-Malea, E., Harzing, A.-W., and López-Cózar, E.D. 2017. "Can We Use Google Scholar to Identify Highly-Cited Documents?," *Journal of Informetrics* (11:1), pp. 152-163.

Page, L., Brin, S., Motwani, R., and Winograd, T. 1999. "The Pagerank Citation Ranking: Bringing Order to the Web," Stanford InfoLab.

Peffers, K., Tuunanen, T., Rothenberger, M.A., and Chatterjee, S. 2007. "A Design Science Research Methodology for Information Systems Research," *Journal of management information systems* (24:3), pp. 45-77.

Shetty, N. 2016. "Query Suggestions to Help Explore New Topics." Retrieved April 21 2017, 2017, from https://scholar.googleblog.com/2016/06/query-suggestions-to-help-explore-new.html

Skiera, B., Eckert, J., and Hinz, O. 2010. "An Analysis of the Importance of the Long Tail in Search Engine Marketing," *Electronic Commerce Research and Applications* (9:6), pp. 488-494.

Vaishnavi, V., and Kuechler, W. 2004. "Design Research in Information Systems,").

Van Deursen, A.J., and Van Dijk, J.A. 2009. "Using the Internet: Skill Related Problems in Users' Online Behavior," *Interacting with computers* (21:5), pp. 393-402.

Verstak, A., Acharya, A., Suzuki, H., Henderson, S., Iakhiaev, M., Lin, C.C.Y., and Shetty, N. 2014. "On the Shoulders of Giants: The Growing Impact of Older Articles," *arXiv preprint arXiv:1411.0275*).

Von Alan, R.H., March, S.T., Park, J., and Ram, S. 2004. "Design Science in Information Systems Research," *MIS quarterly* (28:1), pp. 75-105.