6-2017

A Novel Improvement to Google Scholar Ranking Algorithms Through Broad Topic Search

Matthew Russell Kearl
Dakota State University, kearl@dixie.edu

Cherie Bakker Noteboom
Dakota State University, cherie.noteboom@dsu.edu

Deb Tech
Dakota State University, deb.tech@dsu.edu

A Novel Improvement to Google Scholar Ranking Algorithms Through Broad Topic Search

Matthew Kearl
Dakota State University
kearl@dixie.edu

Cherie Noteboom
Dakota State University
cherie.noteboom@dsu.edu

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 and citation count to sort its results. This paper presents a novel improvement to Google Scholar algorithms by aggregating multiple synonymous searches into one set of results, offsetting the necessity to guess all potential search phrases for a research topic. This design science research method uses a broad topic analysis that examines search queries, finds synonymous phrases, and combines all keyword searches into one set of results based on current Google Scholar citation count algorithms. To support and evaluate this research-in-progress, several users will compare multiple niche search queries against old and new algorithms. The expectation of this design is to introduce modern algorithm techniques to academic search engines, resulting in greater quality, discoverability, and core topic diversity of published research.

Keywords

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

INTRODUCTION

Google Scholar (GS) has been gaining traction as a critical tool for research and discovery of new scientific research papers (SRP). GS uses algorithms to rank and return the most relevant results from a user's search query. Depending on these algorithms, relevant or less relevant results will be returned to the user. When compared to modern web search engines today, GS algorithms are rudimentary and primarily based on citation count and exact keyword match, and often do not provide as relevant results as it might otherwise (Beel and Gipp 2009b).

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 Search Engine Results Page (SERP). This research aims to investigate current algorithm advantages and pitfalls, then proposes a new ranking method to be developed and tested for relevancy 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 save SRP 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 through counting inbound citation count (Beel and Gipp 2009a). 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. 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). 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. The problem with exact match is that it does not allow synonymous keywords to be found. For example, a query for "police abuse" will result in a completely different SERP than, "cop abuse". Both queries mean the same, but if the researcher does not consider synonymous key phrases, high-quality research could be missed.

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. Although it is a good practice to make research accessible and more visible, over-optimized results through keyword stuffing may prevent important works from being found by pushing them lower in the SERP (Beel et al. 2009).

Other researchers have proposed methods to counteract citation rank signals, that include a time depreciation score (Amolochitis et al. 2013), publication venue signal (Hasson et al. 2014), and term frequency heuristics (Amolochitis 2014). Although these novel solutions propose excellent approaches, none address the problem of limited results due to exact keyword match, and the exclusion of potential long tail keyword searches (Dennis 2016).

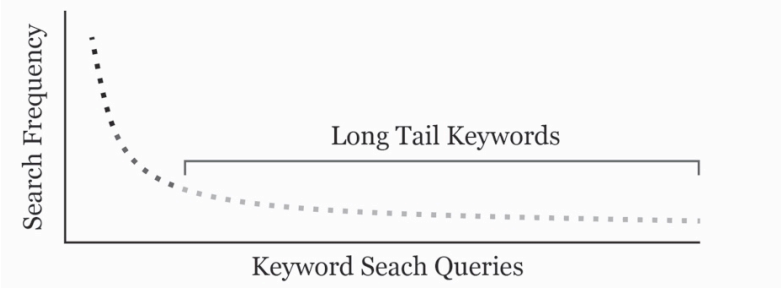

Figure 1 – Long tail keywords (Skiera et al. 2010)

As shown above in figure 1, long tail search terms refer to the uncommon synonymous keywords that a user might not consider when performing a search. Because long tail search terms might not be entered as frequently, those results will receive low visibility on the SERP no matter the quality. Modern web search engines have solved high-quality SRP discoverability issues through broad topic search. 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) by developing a topic-based search on GS that would 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.

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 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 and produce a combined list of results sorted by similar algorithms of GS based on citation count. These algorithms could potentially continue to strengthen the Matthew Effect due to broad topic search, however the purpose of this process is to diversify results based on long tail keywords that researchers may neglect.

If this proposed system is to be effective, in process P5, an improved SERP will be presented to the user and found of higher quality and value.

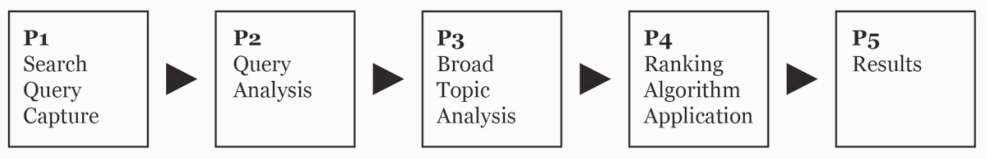

Figure 2 – Broad Topic Analysis Processes

To measure the value and quality of ranked SRP, a group of academic researchers will be given a niche and asked to organically select a search phrase of their choice, then compare the original GS results to the new broad topic SERP. 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.

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 of 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.

REFERENCES

1. Al-Hattab, F. M. F. (2016). "An Efficient Ranking Algorithm for Scientific Research Papers." Zarqa University-Jordan.
2. 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.
3. 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), pp. 1326-1343.
4. 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.
5. 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.
6. Beel, J., and Gipp, B. (2010). "Academic Search Engine Spam and Google Scholar's Resilience against It," *Journal of electronic publishing* (13:3).
7. 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.
8. 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.

9. 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.
10. Dennis, J. (2016). "Search Engine Optimization and the Long Tail of Web Search," in: *Department of Linguistics and Philology*. Uppsala University.
11. 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.
12. Labbé, C. (2010). "Ike Antkare One of the Great Stars in the Scientific Firmament," *International Society for Scientometrics and Informetrics Newsletter* (6:2), pp. 48-52.
13. Marks, T., and Le, A. (2016). "Increasing Article Findability Online: The Four C's of Search Engine Optimization,").
14. 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.
15. 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.
16. Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). "The Pagerank Citation Ranking: Bringing Order to the Web," Stanford InfoLab.
17. 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.
18. 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.
19. 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.