Selected Papers of #AoIR2022:
The 23rd Annual Conference of the
Association of Internet Researchers
Dublin, Ireland / 2-5 Nov 2022

'SORT BY RELEVANCE' – WHOSE RELEVANCE? A CRITICAL EXAMINATION OF ALGORITHM-MEDIATED ACADEMIC LITERATURE SEARCHES

Katy Jordan
University of Cambridge

Introduction and background

The Internet has revolutionised access to information. It has arguably never been easier to access academic literature and research evidence, and a range of online scholarly databases and academic social networking sites provide platforms to search the literature. While this has the potential for a much wider audience – academic and non-academic – to access research literature, there are reasons to be cautious about the platforms which mediate access. For example, different databases vary in terms of their coverage and the sources they include. Furthermore, many are run as commercial enterprises, which may bring hidden priorities for promoting particular content. As a result, relying on a particular database as a source may bring only a partial view of a research field.

The use of algorithms – even if intended to aid the user, by providing what it calculates to be the most 'relevant' material - can further obscure exactly why particular literature has been included in search results. For example, following in the wake of its market dominance as a search engine, Google Scholar is an extremely popular way of searching the academic literature, and uses an algorithm to determine the order in which search results are presented:

> "Google Scholar aims to rank documents the way researchers do, weighing the full text of each document, where it was published, who it was written by, as well as how often and how recently it has been cited in other scholarly literature." (Google Scholar, 2021).

While Google Scholar has received a great deal of research attention in the bibliometric field, studies have more frequently focused upon questions relating to the size and coverage of the database, and reliability of citation counts. A small group of studies

Suggested Citation (APA): Jordan, K. (2022, November). *'Sort by relevance' – whose relevance? A critical examination of algorithm-mediated academic literature searches*. Paper presented at AoIR 2022: The 23rd Annual Conference of the Association of Internet Researchers. Dublin, Ireland: AoIR. Retrieved from http://spir.aoir.org.

have focused on the algorithm to-date, which largely confirm the definition above (Beel et al., 2010), and provide some further insight. Recently, Rovira et al. (2018) identify the number of citations as the major factor influencing ranking; this analysis was extended to other platforms, which showed that this is also the case for Microsoft Academic (now defunct), but not Scopus, and that citations may influence ranking in Web of Science despite not being described as part of their model (Rovira et al., 2019). Furthermore, Rovira et al. (2021) show that the algorithm is biased in favour of articles written in English.

The Google Scholar algorithm description demonstrates potential risks for algorithm-mediated literature searches. First, by drawing upon citations and favouring certain publication outlets, the rankings are likely to amplify the inequalities present in scholarly publishing. Second, it carries methodological risks when carrying out literature reviews, as it is not clear exactly why a particular article has deemed to be of high relevance in search results; some have cautioned against relying on Google Scholar within systematic reviews for example for this reason (Giustini & Kamel Boulos, 2013). However, while awareness of the risks of algorithms is growing, including in relation to academic research and higher education (Matthews, 2021), there is a lack of research at present. This paper reports on a project which is using a mixed-methods approach to explore how academics navigate these risks in practice.

Surveying the prevalence of algorithm in literature searches

While the Google Scholar algorithm has received a degree of research interest and concerns have been raised about the opacity of its rankings, it is no longer an isolated case. The first step in this research project involved examining how prevalent similar ‘sort by relevance’ algorithms currently are in the context of academic bibliographic databases. Fourteen of the largest (general rather than subject specific) platforms were systematically examined. All were found to offer the option to sort by ‘relevance’ or ‘best match’, and this was the default setting in all but two cases. Definitions were provided by just over half of the platforms (eight of the 14 reviewed). The following platforms included sorting by relevance but did not provide definitions: Academia.edu, Aminer, Arxiv, Core.ac.uk, Lens.org, and Zenodo. Those which did provide definitions are listed in Table 1. The level of detail provided in the definitions varied, typically being a short paragraph or bullet points. A notable exception is the Semantic Scholar platform, which publishes technical details of its algorithm through GitHub (Feldman, 2020). The types of information which different platforms state that their algorithms use are mapped in Table 1 (although it is important to caution that little detail was provided, and this may not account for all data used). This reveals that the platforms which do provide information about their ranking algorithms can be broadly divided into two groups: those which are primarily focused on the text included in articles, and those which use other attributes such as the author, where published, date published (favouring recent publications), and the number of citations received.

Table 1: Mapping of types of data used in definitions of ‘relevance’.

Platform	Data sources used				
	Text weighting	*Publication*	*Author*	*Citations*	*Date published*
Dimensions.ai	■				
Google Scholar	■	■	■	■	
JSTOR	■				■
Pubmed	■			■	■
Science Direct	■				
Scopus	■				
Semantic Scholar	■	■	■	■	■
Web of Science	■				

Conclusions

This paper has presented a critical look at the use of algorithms to rank the results of academic literature searches according to the opaque concept of ‘relevance’. The survey of platforms confirms that while the issue of sorting ‘by relevance’ is typically associated with Google Scholar, it is now a common feature across scholarly databases. While a definition of ‘relevance’ is coded into the algorithms themselves, this is rarely presented with clarity – and often not available at all. The combination of data sources which feed such algorithms also carries a risk of exacerbating biases in academic publishing. This opens up a range of further questions about the assumptions academics hold about how algorithms in literature databases work, and the practical implications of this for facilitating open, unbiased access to knowledge. A survey and interviews with academics are planned as the next steps for this research, and emerging findings from these research activities will also be discussed in the session.

Acknowledgments

This research is supported by funding from the Society for Research in Higher Education (SRHE).

References

Beel, J., Gipp, B. & Wilde, E. (2010) Academic search engine optimization (ASEO). Optimizing scholarly literature for Google Scholar & co. *Journal of Scholarly Publishing*, 41, 176–190.

Feldman, S. (2020) Building a better search engine for Semantic Scholar. *AI2 blog.* https://blog.allenai.org/building-a-better-search-engine-for-semantic-scholar-ea23a0b661e7

Giustini, D. & Kamel Boulos, M.N. (2013) Google Scholar is not enough to be used alone for systematic reviews. *Online Journal of Public Health Informatics*, 5(2), 1-9.

Google Scholar (2021) About Google Scholar. https://scholar.google.com/intl/en/scholar/about.html

Matthews, D. (2021) Will a Facebook-style news feed aid discovery or destroy serendipity? *Times Higher Education*. https://www.timeshighereducation.com/features/will-facebook-style-news-feed-aid-discovery-or-destroy-serendipity

Rovira, C., Guerrero-Solé, F. & Codina, L. (2018) Received citations as a main SEO factor of Google Scholar results ranking. *El Profesional de la Informacion*, 27(3), 559–569.

Rovira, C., Codina, L., Guerrero-Solé, F. & Lopezosa, C. (2019) Ranking by relevance and citation counts, a comparative study: Google Scholar, Microsoft Academic, WoS and Scopus. *Future Internet*, 11(9), 202. doi: 10.3390/fi11090202

Rovira, C., Codina, L. & Lopezosa, C. (2021) Language bias in the Google Scholar ranking algorithm. *Future Internet*, 13(2), 31. doi: 10.3390/fi13020031

Yu, K., Mustapha, N. & Oozeer, N. (2017) Google Scholar's filter bubble: An inflated actuality? In: Esposito, A. (Ed.) *Research 2.0 and the impact of digital technologies on scholarly inquiry.* IGI Global, pp.211-229. doi: 10.4018/978-1-5225-0830-4.ch011