157

‘Sort by relevance’: Exploring user assumptions about ranking in online academic literature searches

Katy Jordan, Sally Po Tsai
University of Cambridge, Cambridge, United Kingdom

Research Domains

Digital University and new learning technologies (DU)

Abstract

Access to academic literature has arguably never been easier than it presently is, through the Internet and online databases (such as Google Scholar). However, as the sheer scale of information available online has grown, algorithms are increasingly used in order to deal with searching - and this is also the case in relation to the academic literature. It is now common for online databases to provide search results sorted ‘by relevance’. However, how relevance is defined is not clear, and varies according to different platforms. In this session we will report findings from a recent mixed-methods study undertaken with the goal of understanding academics’ beliefs and assumptions about how such rankings work. Data collection includes an online survey, and in-depth interviews with a sub-sample of participants. The findings of the study will be presented, and the practical implications for academics will be discussed.

Full paper

The academic literature is arguably more accessible than ever before, through the wide array of platforms which act as sources to allow searching and access via the internet. Platforms - such as Google Scholar - which provide access to an ever increasing body of academic literature often utilise ranking algorithms in order to manage how search query results are prioritised and presented to users. The rankings ‘by relevance’ introduce an opaque layer to how academics engage with the literature, with potentially important implications for rigour and equity. In this paper presentation, we will present findings from a study which has been undertaken to explore these issues.

Such algorithms are intended to aid the user, by providing a calculated way to present the most ‘relevant’ material from an unmanageably large number of search results – but can also obscure exactly why particular literature has been included in search results. The risk of receiving a restricted or biased view of a research field when undertaking a literature search is potentially heightened by use of ranking algorithms in the presentation of search results. This lack of transparency could have negative impacts on the rigour of literature reviews and potentially risk creating ‘filter bubbles’ (Matthews, 2021). Furthermore, depending on the types of information used in the ranking algorithms, there is a risk that the way in which results are prioritised may exacerbate existing biases within academic publishing.

A popular and prominent example of a platform which utilises sorting by relevance is Google Scholar. The Google Scholar website describes its ranking as follows:

“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, 2022).

While a body of research literature exists on the topic of Google Scholar, a larger body of research focuses upon the relative size and coverage of its database, while few studies have focused upon how results are ranked. Findings from previous studies provide further detail to the definition above, and confirm that the definition of relevance combines both the content and social information about articles. The date of publication, keywords in the title (Beel & Gipp, 2009) - but not the frequency of keywords in the text of the article (Beel & Gipp, 2009) or synonyms of keywords (Kearl et al., 2017) - have been identified as key factors in determining the ranking. Furthermore, the number of citations received (Rovira et al., 2019), author reputation (Google Scholar, 2022), reputation of the publication or domain (Google Scholar, 2022), and the language in which a document is published (Rovira et al., 2021) are also positioning factors in the Google Scholar relevance ranking algorithms. The latter group of factors draw upon more social information rather than solely the content of articles. For example, numbers of citations and reputation metrics for journals may reflect biases in academic publishing (e.g. Czerniewicz, 2016; Larivière et al, 2013) - and combining such factors may exacerbate this. However, ‘sorting by relevance’ is no longer unique to Google Scholar, and now a typical feature in online academic literature databases - but how relevance is defined is rarely clear, and varies according to platform (Jordan, 2022 forthcoming).

In this paper presentation, we will report findings from a recent study undertaken in order to examine the assumptions that academics’ have about how such ranking algorithms work, and the practical implications search algorithms present for engaging with the research literature. The study uses a mixed methods approach (taking a lead from a recent study focused on users assumptions about the Tik Tok algorithm; Klug et al., 2020), with an initial online survey followed by in depth semi-structured interviews. We will discuss the findings, and their implications for academic practice.

References

Beel, J. & Gipp, B.(2009) Google scholar’s ranking algorithm: An introductory overview. In: Proceedings of the 12th International Conference on Scientometrics and Informetrics, ISSI’09, Istanbul, Turkey, 14–17 July 2009; pp. 230–241.

Czerniewicz, L. (2016) Knowledge inequalities: A marginal view of the digital landscape. Keynote presentation at Open Repositories Conference 2016, Dublin, Ireland, 14 June 2016. Retrieved from https://www.slideshare.net/laura_Cz/laura-czerniewicz-open-repositories-conference-2016-dublin

Google Scholar (2022) About Google Scholar. Retrieved from https://scholar.google.com/intl/en/scholar/about.html

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

Kearl, M. R., Noteboom, C., & Tech, D. (2017). A proposed improvement to Google Scholar algorithms through broad topic search. Proceedings of the Twenty-third Americas Conference on Information Systems, Boston, 2017. Retrieved from https://scholar.dsu.edu/bispapers/6/

Klug, D., Qin, Y., Evans, M. & Kaufman, G. (2020) Trick and please: A mixed-method study on user assumptions about the TikTok algorithm. 13th ACM Web Science Conference 2021 (WebSci’21), Virtual Event, United Kingdom.

Larivière, V., Ni, C., Gingras, Y., Cronin, B. & Sugimoto, C. R. (2013) Bibliometrics: Global gender disparities in science. Nature 504, 211–213).

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

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

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. https://doi.org/10.3390/fi11090202