Making spatial and temporal Wikipedia browsing habits accessible

Andrés Gvirtz,
King's College London

Jascha Achterberg
University of Cambridge

Pauline L. Pfuderer
University of Cambridge

Abstract

This project aims to improve our understanding of Wikipedia's readership, while actively strengthening, expanding, and diversifying the research community using Wikipedia. To this end, we are building a platform that allows interdisciplinary researchers to investigate the spatial and temporal Wikipedia browsing habits over a period of 8 years, significantly improving our understanding of what, where, when and how people are reading. Aiming to involve researchers in the social sciences and humanities, we focus on accessibility, with all analyses and visualisations performed without requiring any coding skills by the end-user. Overall, the project aligns with Wikimedia's 2030 Strategic direction of knowledge as a service and knowledge equity.

Introduction

In June 2023 Wikimedia released a treasure chest of data: The daily page views of every single Wikipedia page by country going back 8 years. This feature was implemented following sustained interest in understanding the readership (1) and several community requests. The reasons for wanting the data are as varied as Wikipedia's community: some are interested in comparing interest in topics across countries, others are interested in zooming into a country, understanding what languages people from that country are reading, and if it influences topics of interest. The questions are diverse, but with the current solution, all require a profound technical understanding and data science skills to process the ~2,000 individual datasets of the latest release. The size and complexity of the dataset is a major barrier towards the Wikimedia 2030 strategic priority of knowledge equity and the wide adoption in research. We therefore suggest an easy-to-use no-code platform to allow observation of spatial and temporal trends in Wikipedia browsing habits.

Date: June 1, 2024 to June 30, 2025.

Related work

Prior research on Wikipedia's readership includes general behaviour analyses and focused studies on topics like medical content (2, 3). However, a recurring observation is that much of this research remains confined within specific academic or thematic silos. Additionally, there is a notable gap in making the data and findings broadly accessible (4). The multilanguage structure is also of interest for researchers outside of classical computer science domains, but challenging to analyse systematically (5). We therefore aim to make understanding the readership across topics, countries and languages findable, accessible, interoperable and reusable.

Methods

The daily data releases of the differential privacy page counts are the input dataset for our research tool. We map all topics across different

languages, allowing to map the interest in topics as diverse as, e.g. Artificial Intelligence and Zoning, across all languages and countries. We further normalise the interest in the topic by the overall Wikipedia usage in that language and country. We have already built a working prototype for this part. Subsequently, the user can set parameters of interest, e.g. timeframe, absolute or relative search term frequency and country or language level requests. The implementation is done in Python, specifically utilising pandas, numpy, urllib, requests, beautifulsoup4 and seaborn for data processing and visualisation. The final product will be wrapped to reduce complexity further.

Expected output

We want to give researchers of all disciplines, as well as journalists and members of the public the ability to get data-driven insights from Wikipedia. We envision this similar to Google Trends, which is used by researchers and the public for a variety of reasons (6). We plan to publish the source code, and the resource in an academic journal. We have earmarked Scientific Data for this, but are open to suggestions. To ensure the dissemination in research circles, we want to share the work at the International Conference on Computational Social Science (ic2s2).

Risks

The overall delivery of the project is modularised, reducing dependencies between work packages and hence increasing feasibility and adherence to project timeline. Most importantly, we further separated the project into three work streams, where the first one standalone already adds value. The first step is to build a working version of all the features outlined. We will do so on Google Collab as a first step, to allow for easy collaboration and testing, and share this with beta testers. The notebooks will already provide the research benefit to all the users in the computational community. Step 2 is to work on a wrapper and a standalone solution that allows access through a visual interface. This will greatly enhance the usability for the social science and humanities research community. Step 3 would be a hosted app solution, with commonly requested queries and search terms already pre-computed to make the experience quicker. This would especially have value for the public.

Community impact plan

We hope to work with the team behind the Wikimedia differential privacy release, looking into long-term hosting at Wikimedia. We believe that Wikipedia data is uniquely positioned for cross-cultural work, considering its reach and consistent standardisations across queries, which are not possible with e.g. most search engine data.

Evaluation

We will have continuous assessment before, during and after the development. We will use interviews to understand what features are most needed, use surveys to understand the usability and ease of access of the MVP and prototypes, monitoring adoption to evaluate impact.

Budget

We are requesting 35.5k, of which 37% will be paid on salaries, specifically hiring JA and PLP as part time research associates for 6 months each. Computational tools and hosting come in at 25%, accounting for temporary solutions, such as Google Collab for the developer team, and subsequent hosting costs accounting for multi-user demands. The planned conference attendance for the team comes in at 18%, including registration, economy flights and

accommodation. 7% is budgeted for the APC charge. Institutional overhead is 13%.

Prior contributions

We are computational researchers from three different areas, with publications in machine learning, psychology, business and biomedical journals. We have previously been involved in developing open-source software (7) and methods (8, 9) in academic and industry settings. The PI on the project is also closely aligned to the computational social science community, having presented e.g. at ic2s2.

References

1) Okoli, C., Mehdi, M., Mesgari, M., Nielsen, F. Å., & Lanamäki, A. (2014). Wikipedia in the eyes of its beholders: A systematic review of scholarly research on Wikipedia readers and readership. *Journal of the Association for Information Science and Technology*, 65(12), 2381-2403.
2) Ahn, B. G., Van Durme, B., & Callison-Burch, C. (2011, June). WikiTopics: What is popular on Wikipedia and why. *In Proceedings of the workshop on automatic summarization for different genres, media, and languages* (pp. 33-40).
3) Heilman, J. M., & West, A. G. (2015). Wikipedia and medicine: quantifying readership, editors, and the significance of natural language. *Journal of Medical Internet Research*, 17(3), e62.
4) Schroeder, R., & Taylor, L. (2015). Big data and Wikipedia research: social science knowledge across disciplinary divides. *Information, Communication & Society*, 18(9), 1039-1056.
5) Torres-Simón, E. (2019). The concept of translation in Wikipedia. *Translation Studies*, 12(3), 273-287.
6) Jun, S. P., Yoo, H. S., & Choi, S. (2018). Ten years of research change using Google Trends: From the perspective of big data utilizations and applications. *Technological forecasting and social change*, 130, 69-87.
7) Jones, M. J., Rai, S. K., Pfuderer, P. L., Bonfim-Melo, A., Pagan, J. K., Clarke, P. R., ... & Boemo, M. A. (2022). A high-resolution, nanopore-based artificial intelligence assay for DNA replication stress in human cancer cells. *bioRxiv*, 2022-09.
8) Achterberg, J., Akarca, D., Strouse, D. J., Duncan, J., & Astle, D. E. (2023). Spatially embedded recurrent neural networks reveal widespread links between structural and functional neuroscience findings. *Nature Machine Intelligence*, 1-13.
9) Eckert, F., Gvirtz, A., Liang, J., & Peters, M. (2020). A method to construct geographical crosswalks with an application to us counties since 1790 (No. w26770). National Bureau of Economic Research.