Geo-cultural influences in Wikipedia Usage: improving access, diversity and relevance in search engines

Javier López Otero—Universidad Isabel I

Ángel Obregón Sierra—Universidad Isabel I

Antonio Gavira Narváez—Universidad Isabel I

Abstract

This project will analyze how the geographical and cultural environments influence in Wikipedia searches to improve accessibility and representativeness of contents. This will require the analysis of search trends and the cultural and linguistic influences using Google Trends and GIS. Data will be gathered through surveys, focus groups, and data mining to be later analyzed through machine learning. The results will include an interactive map of queries, informs on cultural diversity, predictive analysis models, content optimization guidelines, expansion strategies, community participation models, and an analysis of impact in knowledge access, benefiting editors, researchers and Wikimedia.

Table of Contents

1 Introduction

In the digital era, Wikipedia has come to be a key resource for *gratis*, accessible knowledge, widening its reach through mobile devices. However, there is still a lack of understanding on how the geographical and cultural environment of its users influence their Wikipedia search queries. It is then that we seek to analyze how these specific environs influence search patterns in Wikipedia. This information can be used to improve User Experience by favoring access to information better suited to the user's query intention. This analysis can be essential to the Wikimedia Foundation, since understanding its dynamics is essential to keep and improve Wikipedia's relevance and efficiency

2 Timeline

June 1, 2024 to June 1, 2025

3 Related work

There are works by Miquel-Ribé and Laniado (2018, 2019) analyzing the cultural diversity of Wikipedia, and Teplitskiy, Lu, and Duede (2016) analyzed the influence of academic status and accessibility in Wikipedia's references. Other related works are Meeks (2011) on cartography in Wikipedia, Hu et al. (2009) on query intentions, Kousha and Thelwall

(2016) on the academic impact of Wikipedia's citations, and Lewandowski and Spree (2010) on the quality of articles in search engines. However, even though there are similar works, the direct impact of the geographic-cultural context on Wikipedia search queries is understudied still.

4 Methods

The phases are as follows:

1. Data recollection: Analysis of user search queries on search engines: tools such as Google BigQuery, Google Trends and Google Public Data Explorer will be used to extract data on user queries, focusing on location and language of these queries.

    1. Spatial data: the spatial queries will be mapped with GIS, with the aid of Google Maps and Bing Maps. The geographical area of this research will be in 5 sites and 5 countries or regions with different languages.

    2. Non-spatial data: We will analyze the position of Wikipedia in the results of non-spatial queries in several languages.

    3. Surveys and focus groups: Survey design: surveys for active and inactive Wikipedia users, studying their habits and search preferences.

    Focus groups will be organized to understand search preferences and behavior.

2. Data analysis:

    1. Pattern analysis in queries: data mining and statistical analysis to examine patterns in spatial and temporal queries, taking into account variables such as geographical location, language and subject. Development of machine learning [-powered] predictive models to understand the relationships between location, cultural and linguistic context, and queries. These models will include query data, user perceptions and internal metrics from Wikipedia articles.

5 Expected results

Creation of an interactive map with Wikipedia usage trends. It will identify the geographical and linguistic areas that need more attention.

Creation of predictive analysis models to aid editors to anticipate and respond to emerging information needs.

Content optimization guidelines according to search trends will increase visibility and accessibility of content.

The target audience of this proposal are Wikipedia editors, admins, and users, as well as people in academia.

6 Risks

Assembling a representative sample of participants for the focus groups can be a challenge. Implementation of improvements in Wikipedia can be met with resistance or practical challenges, which would limit this project's impact.

7 Impact plan

Maximization of this project's impact will be coherent with the 2030 Movement Strategy. Additionally, we will carry out other actions such as: collaboration with Wikimedia communities, participation in Wikimedia events, publishing of results in academic and communication formats.

8 Evaluation

Evaluating this project about Wikipedia queries according to their geographical and cultural environment requires clear and measurable success criteria. It includes evaluating the efficiency of data recollection and analysis in order to find search patterns, effectiveness of surveys and focus groups in terms of quality and representativeness of data, precision and robustness of the predictive models, and the relevance and utility of these results for the Wikimedia community.

9 Budget

Wages for three part-time technicians, specialized in analysis of statistical, spatial and informational data; 30,000 €

Cost of publishing in open access journals; 4,000 €

General institutional expenses; 3,400 €

Total requested: 37,400 €

10 Main contributions

Ángel Obregón is an admin in the Spanish Wikipedia (User:Vanbasten_23). He has contributed to Wikimedia Foundation projects for 17 years, mostly in technical spaces, programming bots in several languages, and Wikidata. He has published 14 scientific papers about Wikipedia, mostly dealing with education, and other 4 about Wikidata.

Javier López is a lecturer at the Isabel I University, has published over 20 papers and book chapters, has been part of 5 I+D+I projects[1] in Spain and has lectured in over 10 events abroad.

Antonio Gavira is a lecturer at the Isabel I University, has published over 30 papers, book chapters and communication materials of several congresses. Has also been part of several I+D+I projects in technological subjects.

The most recent research experience in Wikipedia would be *«Methodology for the Incorporation of Geographic Information in Wikidata»* and *«Unveiling Wikipedia's Role in Urban Tourism: An In-depth Analysis of Destination Choice Using Artificial Intelligence Indicators»* (still under evaluation)

Annex 1: Methodology

Phase 1: Data recollection

- Analysis of User search queries in search engines
- Recollection of spatial data
- Recollection of non-spatial data
- Surveys and focus groups
- Factors explaining the position of web pages in browsers

Phase 2: Data analysis

- Analysis of trends in spatial and temporal queries
- Development of predictive models
- Validation and adjustment of models

References

Hu, Jian, Gang Wang, Fred Lochovsky, Jian-tao Sun, and Zheng Chen. 2009. "Understanding User's Query Intent with Wikipedia." In *Proceedings of the 18th International Conference on World Wide Web*. WWW '09. ACM. https://doi.org/10.1145/1526709.1526773.

Kousha, Kayvan, and Mike Thelwall. 2016. "Are Wikipedia Citations Important Evidence of the Impact of Scholarly Articles and Books?" *Journal of the Association for Information Science and Technology* 68 (3): 762–79. https://doi.org/10.1002/asi.23694.

Lewandowski, Dirk, and Ulrike Spree. 2010. "Ranking of Wikipedia Articles in Search Engines Revisited: Fair Ranking for Reasonable Quality?" *Journal of the American Society for Information Science and Technology* 62 (1): 117–32. https://doi.org/10.1002/asi.21423.

[1] *N.T.:* Research, Development and Innovation. This particular nomenclature is sometimes used for government-led projects in Spain.

Lewoniewski, Włodzimierz, Ralf-Christian Härting, Krzysztof Węcel, Christopher Reichstein, and Witold Abramowicz. 2018. “Correction to: Application of SEO Metrics to Determine the Quality of Wikipedia Articles and Their Sources.” In *Information and Software Technologies*, E1–1. Springer International Publishing. https://doi.org/10.1007/978-3-319-99972-2_49.

López-Otero, J., A. Gaviran-Narvaez, and Vega-Pozuelo R.F. 2024. “In Depth Analysis of Destination Choice Using Artificial Intelligence Indicators.” *Social Indicators Research*.

Meeks, Elijah. 2011. “Mapping Wikipedia: Geolocated Articles as a Proxy of Culture and Attention.” Online. https://dhs.stanford.edu/spatial-humanities/mapping-wikipedia-geolocated-articles-as-a-proxy-of-culture-and-attention.

Miquel-Ribé, Marc, and David Laniado. 2018. “Wikipedia Culture Gap: Quantifying Content Imbalances Across 40 Language Editions.” *Frontiers in Physics* 6 (June). https://doi.org/10.3389/fphy.2018.00054.

———. 2019. “Wikipedia Cultural Diversity Dataset: A Complete Cartography for 300 Language Editions.” *Proceedings of the International AAAI Conference on Web and Social Media* 13 (July): 620–29. https://doi.org/10.1609/icwsm.v13i01.3260.

Obregón-Sierra, Ángel, Javier López-Otero, Antonio Gavira-Narváez, and Rafael Vega-Pozuelo. 2023. “Methodology for the Incorporation of Geographic Information in Wikidata.” *Revista de Estudios Andaluces*, no. 45. https://doi.org/10.12795/rea.2023.i45.

Teplitskiy, Misha, Grace Lu, and Eamon Duede. 2016. “Amplifying the Impact of Open Access: Wikipedia and the Diffusion of Science.” *Journal of the Association for Information Science and Technology* 68 (9): 2116–27. https://doi.org/10.1002/asi.23687.