ARISE-Wiki: Alternative Routes for Informed Exploration of Wikipedia

Michele Tizzani
ISI Foundation, Torino, Italy

Anna Sapienza
Università del Piemonte Orientale "A. Avogadro", Alessandria, Italy

Abstract

This project aims to map the relations between external events, content reliability, and user navigation on Wikipedia. The final goal is designing a method that, taking into account how these factors collectively influence information-seeking patterns, provides insights and alternative navigation routes to enhance the user experience while maintaining a focus on critical discernment.

Introduction

Finding information and keeping up-to-date with news and events are among the top three reasons for using the Internet [1, p. 202]. In the search for information, Wikipedia links are extremely common and appear in 67-84% of desktop search engine result pages [2]. Despite most searches ending in a single page load, Wikipedia readers also exhibit navigation routes characterized by long explorations.

Wikimedia researchers previously highlighted that such routes - a.k.a. rabbit holes - tend to be initiated by very popular pages, associated with events that received substantial public attention at the time of search [3].
This stresses the significant impact of news media on information-seeking, yet other factors - type of event, reliability of content, controversiality of topics, etc. - are at play.

A comprehensive understanding of how these factors collectively shape readers' navigation behaviors is key to fostering a better user experience, emphasizing curiosity and critical discernment. We propose to explore these navigational heterogeneities by tackling the following questions:

1. **RQ1:** How does exposure to exogenous events influence information-seeking on Wikipedia?
2. **RQ2:** Do these patterns arise differently when readers are presented with controversial topics within Wikipedia?
3. **RQ3:** Can we recommend alternative routes to readers, while limiting exposure to controversial sources?

Answering these questions will benefit Wikipedia:

1. **Readers**, by suggesting alternative routes to more reliable pages;
2. **Editors**, by providing insights on external events having the potential to impact editing and content integrity;
3. **Researchers**, by extending the work on page trustworthiness and navigation pathways [3] and benefiting the Wikimedia 2030 strategy in the areas of *Emerging platforms and Misinformation.*

Related work

News and social media spotlight major events, such as epidemics, and not only influence public perception but also trigger information-seeking within Wikipedia [4]–[6]. This triggers readers' navigation in Wikipedia [7], but the extent to which this happens and the role of news have not been quantified yet. Our project bridges this gap by mapping the

interconnections between external events and the probability for readers of ending in a “rabbit hole”.
The way readers consume content can be influenced by various factors, including its reliability [8], [9]. To investigate the effect of content types on exploration in Wikipedia, we will build on existing tools detecting characteristics of pages, such as their controversy level [10].

Methods

Each RQ has a related Work Package (WP):

WP1 aims at how exogenous events, as spotlighted by news media, steer readers' navigation patterns on Wikipedia. We will use Wikipedia's pageviews, edits, and clickstreams, complemented by news aggregators (GDELT) and Google searches (Google Trends) data. Using event analysis and NLP models, we will measure the influence of exogenous events on readers' navigation patterns.

WP2 focuses on understanding how readers' navigation patterns manifest when confronted with varying levels of source reliability within Wikipedia. By leveraging Wikipedia's controversy score and clickstream data, we will employ state-of-the-art machine learning techniques to discern the likelihood of readers descending into "rabbit holes" depending on content reliability and their navigation pattern.

WP3 leverages insights from the aforementioned steps to develop a recommender system designed to propose alternative pathways for navigation to Wikipedia readers. The final result will be a model that, guided by collective navigation patterns, provides recommendations to readers of alternative and reliable content.

Expected output

The main output	(1)
Who benefits from it	(2)

WP1

(1) Quantifiable measure of the impact (i.e. effect size) of external events on Wikipedia readers' navigation patterns.
(2) Insights for informed decision-making for Wikipedia editors to understand and respond to user behavior influenced by external events.

WP2

(1) Prediction probability and feature importance of readers descending into "rabbit holes" and their relation to content reliability and navigation patterns.
(2) Insights for readers for informed content consumption, highlighting the risk of controversial sources and guiding them toward alternative content.

WP3

(1) Development of a Recommender System, designed to propose alternative navigation routes for Wikipedia readers.
(2) Insights for readers complementing the one in WP2 and adding recommendations for alternative routes.

Scientific Outlets:

- Dissemination of findings through scientific publications in reputable journals.
- Participation in scientific conferences such as CCS, IC2S2, and WikiWorkshop to engage with the academic and Wikimedia community.

Practical applications:
Our research outputs could be further developed as part of Wikipedia tools in the future. We envision these possible applications:

- A dashboard to explore the interaction between external events and Wikipedia.
- A plugin informing Wikipedia readers of controversial routes and sources and providing alternative information sources.

Risks

Risk	Description	Contingency Plan
Limited access	Access to Wikimedia granular data would significantly enhance our outputs	If further access cannot be granted, our project can rely on Wikimedia's official dumps
Data availability	We rely on GDELT for news traffic data	If discontinued, we can rely on other paid APIs that offer similar services [11], [12]
Method integration	We rely on controversy scores, a research approach under development	If integration challenges arise, we can build scores based on page attributes

Community impact plan

Public Outreach.

- Participation in public venues, such as Research Nights, to communicate our discoveries to a non-technical audience.
- Presence on social media, institutional pages, and creation of a website.

Wikimedia Community. Contribution to:

- A trusted environment for sustainable knowledge sharing and collaboration.
- A more resilient and interconnected Wikimedia community.
- The Wikimedia Movement's 2030 strategic direction.

Evaluation

- Publications and academic attention
- Usable outputs for Wikipedia editors and readers community.
- Open-source models and datasets for replication/extension for Wikimedia researchers.

Budget

Detail	USD
Salary	25000
Equipment	10000
Software	500
Publishing	3000
Overhead	6750
Other	3500
Total	48750

Prior contributions

The PIs Anna Sapienza and Michele Tizzani have extensive experience in modeling digital behaviors [13]–[15] and information-seeking patterns [4], [5].

References

[1] «Digital 2022: Motivations for Using the Internet — DataReportal – Global Digital Insights». Consultato: 7 dicembre 2023. [Online]. Disponibile su: https://datareportal.com/reports/digital-2022-motivations-for-using-the-internet

[2] N. Vincent e B. Hecht, «A Deeper Investigation of the Importance of Wikipedia Links to Search Engine Results», *Proc. ACM Hum.-Comput. Interact.*, vol. 5, fasc. CSCW1, p. 4:1-4:15, apr. 2021, doi: 10.1145/3449078.

[3] «WiSCoM — Wikipedia Source Controversiality Metrics | MisinfoCon». Consultato: 13 dicembre 2023. [Online]. Disponibile su: https://misinfocon.com/wiscom-wikipedia-source-controversiality-metrics-f520f6c02423

[4] N. Gozzi *et al.*, «Collective Response to Media Coverage of the COVID-19 Pandemic on Reddit and Wikipedia: Mixed-Methods Analysis», *J. Med. Internet Res.*, vol. 22, fasc. 10, 2020, doi: 10.2196/21597.

[5] M. Tizzani *et al.*, «Integrating digital and field surveillance as complementary efforts to manage epidemic diseases of livestock: African swine fever as a case study», *PLOS ONE*, vol. 16, fasc. 12, p. e0252972, dic. 2021, doi: 10.1371/journal.pone.0252972.

[6] M. Tizzoni, A. Panisson, D. Paolotti, e C. Cattuto, «The impact of news exposure on collective attention in the United States during the 2016 Zika epidemic», *PLOS Comput. Biol.*, vol. 16, fasc. 3, p. e1007633, mar. 2020, doi: 10.1371/journal.pcbi.1007633.

[7] T. Piccardi, M. Gerlach, e R. West, «Going Down the Rabbit Hole: Characterizing the Long Tail of Wikipedia Reading Sessions», in *Companion Proceedings of the Web Conference 2022*, in WWW '22. New York, NY, USA: Association for Computing Machinery, ago. 2022, pp. 1324–1330. doi: 10.1145/3487553.3524930.

[8] M. D. Vicario *et al.*, «The spreading of misinformation online», doi: 10.1073/pnas.1517441113.

[9] S. Vosoughi, D. Roy, e S. Aral, «The spread of true and false news online», *Science*, vol. 359, fasc. 6380, pp. 1146–1151, mar. 2018, doi: 10.1126/science.aap9559.

[10] E. Borra *et al.*, «Societal Controversies in Wikipedia Articles», in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, in CHI '15. New York, NY, USA: Association for Computing Machinery, apr. 2015, pp. 193–196. doi: 10.1145/2702123.2702436.

[11] News API Search News and Blog Articles on the Web:https://newsapi.org/

[12] NewsCatcher News API https://www.newscatcherapi.com/

[13] Recommending Teammates with Deep Neural Networks | Proceedings of the 29th on Hypertext and Social Media\uc0\u187{}. https://dl.acm.org/doi/abs/10.1145/3209542.3209569

[14] «Exposure to urban and rural contexts shapes smartphone usage behavior | PNAS Nexus | Oxford Academic». Consultato: 14 dicembre 2023. [Online]. Disponibile su: https://academic.oup.com/pnasnexus/article/2/11/pgad357/7442564

[15] A. Sapienza, A. Bessi, e E. Ferrara, «Non-Negative Tensor Factorization for Human Behavioral Pattern Mining in Online Games», *Information*, vol. 9, fasc. 3, Art. fasc. 3, mar. 2018, doi: 10.3390/info9030066.