9^{th} International Conference on Computational Social Science IC^2S^2
July 17-20, 2023, Copenhagen, Denmark

Going Down the Rabbit Hole: Characterizing the Long Tail of Wikipedia Reading Sessions

Keywords: Wikipedia, web navigation, server logs, information needs

Extended Abstract

If you ever visited Wikipedia for simple fact-checking and then, in a few minutes, found yourself learning that *life does not evolve wheels*[1] and that *unsold video games may end up buried in the desert*[2], there is a high chance that you fell into a "wiki rabbit hole", a term inspired by Lewis Carroll's novel *Alice's Adventures in Wonderland,* where the main character reaches an astonishing world by following a white rabbit deep into its burrow.

Similarly, a wiki rabbit hole[3], sometimes called a wiki hole or wiki black hole, is a popular concept in Internet culture often described as a long navigation session where readers, following multiple links, get lost in Wikipedia and learn about a diverse set of topics[4]. The reason that motivates readers to engage in long navigation sessions, with jumps across different topics, is often associated with boredom [2] and curiosity. Given the substantial time we spend consuming online content, understanding the dynamics of how we seek knowledge can offer useful insight into our information needs and support the design of better systems centered around users' interests. Previous work provided an overall characterization of reading sessions on Wikipedia [1]. However, reading sessions are typically short such that the population-wide average does not adequately capture the behavior contained in the long tail. Therefore, this study serves as a complimentary analysis of how readers browse Wikipedia by focusing on long reading sessions associated with rabbit hole navigation.

Data. We use server logs collected over four weeks in March 2021 from English Wikipedia to describe the exploration dynamics and how readers navigate during the long reading sessions. We characterize the long tail of these navigation traces by focusing on sessions with paths composed of at least ten pageviews (i.e., generated by at least nine sequential internal clicks). Given the long-tailed nature of the tree size distribution, this filter leaves us with 216M pageviews aggregated in 8.97M sessions—0.24% of the original navigation sessions.

Summary of findings. By investigating the frequent paths taken by readers falling into a wiki rabbit hole, we observed that characteristics of the article layout are associated with deep navigation trees. A manual inspection of the top 1000 articles that serve most frequently as an entry point for the rabbit hole reveals that many articles refer to recurrent events with seasonal repetitions, such as elections (i.e. 1946 DUTCH GENERAL ELECTION), sports events (i.e. 1979 NBA ALL-STAR GAME), and award ceremonies (i.e. 2019 ACADEMY AWARDS). In total, 68.4% of the articles in this list contain the word "election" in their title, and 87.3% contain a four digits year. When considering the entire dataset, the pages about elections are 0.5% of all the entry-points articles and cover 3.4% of the sessions. The presence of navigational links in the infobox (Fig. 1) to transition between different instances of the same recurrent event supports the type of browsing similar to reading a slideshow.

[1] https://en.wikipedia.org/wiki/Rotating_locomotion_in_living_systems
[2] https://en.wikipedia.org/wiki/Atari_video_game_burial
[3] https://en.wikipedia.org/wiki/Wiki_rabbit_hole
[4] https://xkcd.com/214/

The dynamics of falling into a wiki rabbit hole show differences across the time of the day, the device used, and the topic of the first article. The fraction of sessions with deep trees is overall higher on desktop than mobile devices and increases in both cases at night. Confirming popular belief and previous findings on more general behavior [1], articles about entertainment, sport, politics, and history are more common as starting points for rabbit hole sessions.

Often the navigation of readers falling into the rabbit hole is imagined as a long session that brings the users to a random page of Wikipedia. We verify this assumption by comparing the readers' trajectories in the topics space with a null model obtained from an unbiased random walker. We are interested in observing how the trajectories diffuse in the space with respect to the origin and if the long navigation paths converge, in multiple steps, to random clicking behavior. To compare the two sets of trajectories (human-generated vs. random walker), we proceed in two steps: first, we create a matched dataset by running for each of the 8.9M readers-generated paths a random walk that, starting from the same article, generates a sequence of the same length. For each step, the next article is selected randomly from the list of links available on the page. Then, we assign the respective ORES[5] topics vectors to each article visited in all trajectories.

Overall, the comparison shows that readers tend to stay semantically close to the first page, even for long sessions. Fig. 2b shows the PCA representation of topic vectors of a random sample of trajectories where the first page is centered at the origin. The marginal density distribution shows that the user-generated paths have a higher concentration close to the first page loaded when compared to the larger spread of random exploration. This intuition is reinforced by computing the Mean Squared Displacement (MSD), a metric typically used in physics to measure the dispersion of a particle from the starting position. Fig. 2c shows that the dispersion coefficients of the readers' navigation, stratified for different tree sizes and positions in the path, are almost half compared to the diffusion of a random walk. These observations suggest that readers stay semantically close to the origin –compared to a random walk– even for very long sessions, making the divergence of a rabbit hole session to a random page more an exception than the norm.

Conclusion. In conclusion, the characteristics of these sessions differ from the majority of very short sessions [1], suggesting that rabbit hole sessions satisfy succinctly distinct needs of readers. This work thus provides new quantitative insights into how Wikipedia is used by readers, which could empower the community to make informed decisions around the organization of Wikipedia's content. More generally, we hope to inspire future research on online knowledge consumption and add a small piece to our understanding of Wikipedia readership.

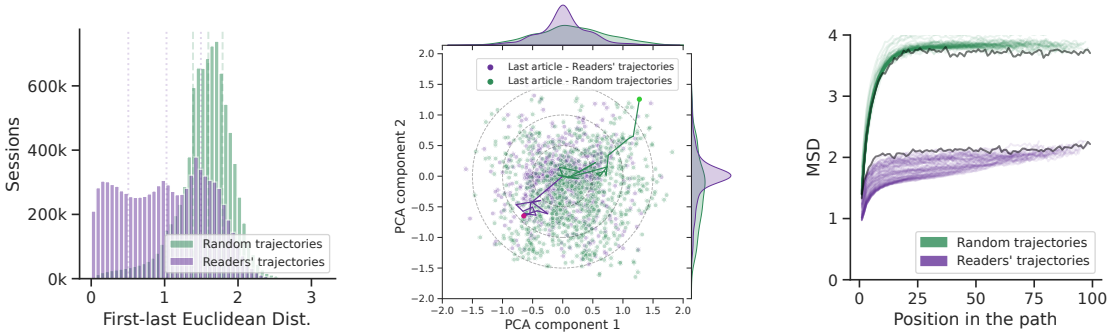
References

[1] Tiziano Piccardi, Martin Gerlach, Akhil Arora, and Robert West. A large-scale characterization of how readers browse wikipedia. *ACM Trans. Web*, jan 2023.

[2] Philipp Singer, Florian Lemmerich, Robert West, Leila Zia, Ellery Wulczyn, Markus Strohmaier, and Jure Leskovec. Why we read Wikipedia. In *Proc. International World Wide Web Conference (WWW)*, 2017.

[5] `https://www.mediawiki.org/wiki/ORES`

Figure 1: Navigational links in the infobox are often used to engage in long sessions.



(a) First-to-last distance (b) PCA projection of navigation trajectories (c) Mean squared displacement

Figure 2: Sessions diffusion in topics space. (2a): Distribution of Euclidean distances between the first and the last article of the path. Quartiles as vertical lines. (2b): First 2 principal components of the last article of 2000 paths generated by human readers and a random walker in the topic space defined by ORES. The first article of the session centered on the origin. Marginal plots show KDE distributions. The green and purple lines represent two examples of full trajectories. (2c): Mean squared displacement (MSD) of the sessions in the topic space defined by ORES. Each line represents the MSD of all the sessions of one specific length [10-100]. The dark trajectories are added for readability and represent the sessions with 100 pageviews.