A Large-Scale Characterization of How Readers Browse Wikipedia

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

Extended Abstract

Given the central importance of information seeking to human nature, understanding how humans seek information and engage with knowledge is of key significance across disciplines, both in the basic and applied sciences. In this work, we provide a perspective in the context of encyclopedic information seeking—an important special case of human information seeking—by leveraging a large-scale dataset of digital traces compiled from one month's worth of English Wikipedia's complete server logs, which offer unprecedented opportunities for observing humans interacting with knowledge in great detail. Wikipedia is a primary source of encyclopedic knowledge and plays a unique role in the global knowledge ecosystem, fulfilling a wide range of information needs. This work is the first to employ the logs in a principled, broad analysis with the goal of systematically elucidating the nature and structure of encyclopedic information-seeking pathways. In other words, whereas previous, complementary research has focused on *why* people use Wikipedia [2], in this paper, we describe the mechanisms of content consumption to investigate the question of *how* people use Wikipedia.

Data. We analyze the server logs of the English edition collected for four weeks between 1 and 28 March 2021. We preprocess the data by removing sensitive information and requests from large networks sharing the same IP. Each request entry includes an anonymous user identifier, the page's title, the timestamp, the access method, the request's referrer, and the relative topics vectors obtained from WikiPDA [1]. The final dataset contains 6.52B pageloads associated with 1.47B user identifiers that we aggregate into navigation trees to model the user sessions. These trees describe how readers traverse Wikipedia by following internal links. We generate a tree by connecting pageloads of the same user via the referrer contained in HTTP headers. Pages reached through internal transitions (i.e. using internal links) are added as children of the most recent load of the article in the referrer, while pageloads with external or *Main_Page* referrers generate a new tree.

Summary of findings. By investigating the root of the navigation trees to understand how readers reach Wikipedia, we found that the most common origin is search engines, representing the source of 77.5% of the incoming traffic. The second most common origin (20%) is the *unknown origin*—or an empty string (i.e., potentially caused by revising the page from browser history). Finally, external websites, including social media platforms, account for 0.95% of the incoming views. The remaining views originate from mobile web views or custom apps.

We observe that the sessions tend to be very short, with 78% of the navigation trees composed of a single pageload (Fig. 1a). However, on average, evening and night sessions are longer than during working hours (Fig. 1b). We investigate further the properties of single-page sessions by fitting a logistic regression to predict if the reader will continue after loading the first article. We represent each first pageload with its topic probabilities (obtained from ORES[1]), device type, and time of day, and obtaining a model with an AUC of 0.606. The coefficients of the regression (Fig. 1b) indicate that longer [shorter] sessions are associated with

[1] https://www.mediawiki.org/wiki/ORES

topical content around entertainment [STEM and medicine]. Additionally, we observe that readers who navigate beyond the first page prefer to navigate in a chain-like fashion without branching (Fig. 2), and by analyzing sequential visits of the same user, we observe that readers frequently–even when two articles are connected–do not use internal links to transition between the two pages but prefer to go to the search engine and re-enter Wikipedia from the SERP. This behavior creates shallow sessions compatible with the random surfer model.

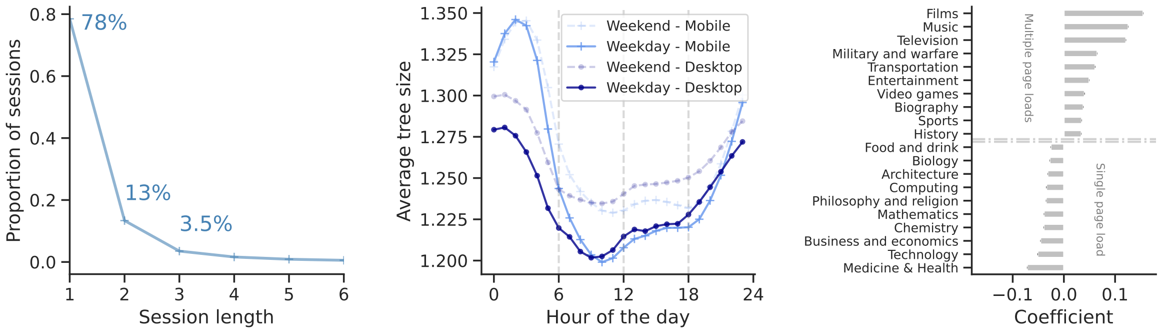
Finally, to shed light on navigation dynamics, we track the evolution of the sessions. Our analysis revolves around three domains: topic space, quality, and network centrality. Here, a navigation tree is represented by the linear path from the root to the last leaf, from where the reader ceases to click further via internal links. In order to better interpret our observations, we compare them with a null model represented by a random walker that navigates by selecting a random link on the page. Investigating how readers diffuse in topic space starting from the first article, we observe that readers move semantically further from the entry point with every step (Fig. 3a) but at a lower rate than the random walker. By tracking how the topical distance to the previous article evolves, we observe a U-shape, suggesting that readers tend to first reduce their semantic step size, before diverging and finally abandoning (Fig. 3b). The evolution of article quality shows a sharp drop at the beginning (Fig. 3c). This behavior can be interpreted as a form of regression to the mean since many sessions start from popular pages with high quality, which thus contribute more to the distribution. Sessions show a sharp drop in quality with the last pageload, indicating that readers have a higher chance to stop Wikipedia-internal navigation when reaching a low-quality page and, as a result, continue navigating in a different branch of the tree or via an external transition. Centrality-based metrics such as out-degree (Fig. 3d) and PageRank (Fig. 3e), correlated with the quality of the article, show similar patterns with a sharp drop with the first step and a second one just before abandoning the session.

Discussion. Our empirical results contribute to describing the relation between contents and navigation, expanding the prior understanding of how the features of the page influence readership and popularity. The navigation of readers on Wikipedia differs from targeted navigation in lab-based settings. We do not observe typical strategies characterized by, e.g., navigation via hubs or gradually decreasing steps in semantic space towards the target. Instead, we find a range of other patterns, such as a U-shape for the step size in semantic space and an immediate sharp drop followed by largely constant centrality measures. Additionally, our findings suggest that upon encountering low-quality pages, readers tend to stop navigating along a specific branch in the navigation tree (and continue along a different branch or stop altogether). This finding is aligned with the definition of information scent used in information foraging theory, stating that readers follow the scent with higher chances of leading to the desired content; when the scent loses intensity, they move to more promising information sources. In light of these observations, our work can have implications for developing theoretical frameworks to describe navigation patterns. Understanding how readers follow specific trails can inform researchers about the properties that guide our search for information and can be instrumental in developing novel theories on how humans move in information networks.

References

[1] Tiziano Piccardi and Robert West. Crosslingual topic modeling with wikipda. 2021.

[2] Philipp Singer, Florian Lemmerich, Robert West, Leila Zia, Ellery Wulczyn, Markus Strohmaier, and Jure Leskovec. Why we read Wikipedia. In *Proc. of WWW*, 2017.



(a) Session length (b) By time (c) Regression coeff.

Figure 1: Navigation trees statistics.

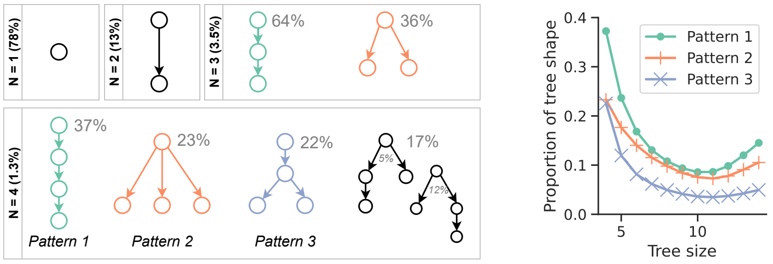
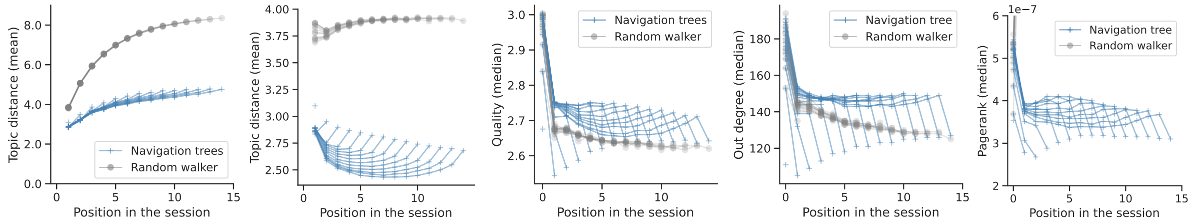


Figure 2: Shape of navigation trees. Frequency of patterns for trees size $N <= 4$ (left panel). Dominance of top three patterns for larger trees (right panel).



(a) First article (b) Prev. article (c) Quality (d) Out-degree (e) Pagerank

Figure 3: Within-session evolution of 5 article properties. Each curve represents sessions of different lengths.