Understanding Search Behavior Bias in Wikipedia*

Bruno Scarone
Northeastern University, USA

Ricardo Baeza-Yates
Northeastern University, USA

Erik Bernhardson
Wikimedia Foundation, USA

Abstract

We analyze Wikipedia's server logs containing millions of search sessions to understand its website search and compare it to generic web search. We segment the studied quantities by the type of client being used and analyze whether there is a relation between the time a user spends on a page and its length, as well as between the time it takes a user to click on a result and the position of the selected page in the ranked list. We rediscover known results in the context of web search that also apply to website search, along with new results in the context of website search.

Keywords: Search behavior, Search biases, Website search, Web dynamics, Wikipedia.

Introduction

Wikipedia is the world's largest and most widely used open encyclopedia on the Web, providing content to millions of readers daily from across the globe in more than 300 actively edited languages, being the 7th most visited site on the Web according to Similarweb.[1] Considering that in the top-6 websites of the previous ranking there are two search engines and three social networks, we could say that Wikipedia is the second most visited content-oriented website, as social networks include content, but it is not their main goal. For this reason, understanding how people search using its website search tool, provides insights into several dimensions including language and devices used, as well as time spent on the platform.

Hence, we focus on three research questions (RQ) to understand the search behavior of Wikipedia users:

- **RQ1**: How is the search behavior affected (biased) by the client type used (desktop vs mobile interface)?
- **RQ2**: What is the relation between the time a user spends on a page and its length, *i.e.*, does page length bias the time a user spends on the page?
- **RQ3**: What is the relation between the time it takes a user to click on a result and the position of the result in the ranking, *i.e.*, does the ranking of a result bias the time it takes a user to click on the link to it?

Data and Methods

All datasets used in this paper consist of data from the server logs of the Wikimedia Foundation. The different data sources used in the analysis are listed in Table 1, together with their description pages as well as the data retention policy that applies to them.

The main data source used is the `dqcd` table, which contains sessionized click-throughs for full searches (*i.e.*, non auto-complete) made by users (*i.e.*, non-bots), for both, mobile and desktop access. The table only contains searches that have a non empty set of click-throughs, *i.e.*, searches that have at least one click, and considers all Wikipedias in the available languages.[2] The top-10 most frequent languages that account for 89% of the total traffic are shown in Table 2. The Search logs table[A] is used to reconstruct the access method in `dqcd`, which was not directly available. The size of pages was obtained from the `wmf.mediawiki_wikitext_current`[B] table.

The quantities analyzed have been computed for 2 one-week time ranges. All time ranges span the first complete week (from Monday to Sunday) of a month. The client type characterization used is the one given in the web request logs' description page[C] by the `access_method` field. We do not include the Wikimedia mobile app in this analysis because its traffic was significantly smaller than the two main methods.

When one of the data sources did not include the `access_method` field, it was reconstructed using the `getAccessMethod` function, implemented in the refinery-source code repository.[D]. Finally, we also used the hourly page views table.[E] When evaluating linear correlation between two variables, we compute Pearson correlation coefficients together with their p-values (< 0.001 in all cases).

*This work was partially funded by NSF Grant 1956096 and the Wikimedia Foundation.

[1] https://www.similarweb.com/top-websites/

[2] The full list can be found at https://meta.wikimedia.org/wiki/Special:SiteMatrix.

Table name	Retention
Search logs	90 days
`discovery.query_clicks_daily` (`dqcd`)	90 days
`wmf.mediawiki_wikitext_current`	-
Hourly views (`wmf.pageview_hourly`)	-

Table 1: Data sources used for this study.

Language	Percentage
English	52.7
German	10.6
Japanese	4.9
Russian	4.2
French	4.1

Language	Percentage
Spanish	3.6
Chinese	2.9
Italian	2.9
Persian	1.8
Polish	1.3

Table 2: Language distribution (top-10) `dqcd` table.

Results

We discuss the main results obtained as part of the analysis when aggregating the data by client (*i.e.*, platform) type. We report the following findings[3]:

RQ1: No significant differences are observed in the quantities studied when segmenting them by client type used (Figure 1), which is not the case in web search.

The distribution of dwell times, the time spent in the answer page, has two modes (Figure 1), with the first one in less than a second and the second in around 7 seconds. The former has been reported previously for mobile ads (Tolomei et al., 2019; Kaplan et al., 2021) as *accidental clicks*. As also happens in search, this should be a generic phenomenon when interacting with a finger on a small screen. What is surprising in this context is to observe the same behavior occurring for desktop too.

RQ2: The time a user spends on Wikipedia pages is in general independent of the size of the answer link clicked, irrespective of the client being used, for dwell times larger than a minute (Figure 2). This behavior was known for a small sample of 25 users (Weinreich et al., 2008), but in this work we use logs containing millions of sessions.

RQ3: We found that the ranking position a user clicks on has no significant impact on the time it takes the user to perform the first click (Figure 3). Given that users rarely view more than 10 results (Jansen and Spink, 2005) (in particular, not the entire list of results), we hypothesize they users may skip results without really reading them, otherwise the time should be linear.

Conclusions and Future Work

We have presented new results regarding the search behavior of Wikipedia users for the two main client types, desktop and mobile. Additionally, we have rediscovered known results in other contexts to be also valid for website search, along with new results in this context. Our contributions categorized across these dimensions are listed in Table 3.

Future work includes providing further insights into what factors produce the observed behavior. Given the in general transient nature of the search behavior, it would be interesting to explore either measures that encourage a more in-depth reading of the articles (and thus longer page stay times) or alternatively others which ensure that the essential information of the articles is located at the beginning of the pages. It would also be valuable to segment users further and see if other behavioral patterns emerge from the data. One can for example evaluate if the search behavior differs across languages, initially analyzing the results for the "average user" for each language. Additional research directions include extending our new results to generic web search and exploring new research questions that should also address the semantics of the queries being issued. This includes most popular languages, most popular contents, and queries that are not well satisfied by the underlying system.

[3]Further details on our results can be found in (Scarone et al., 2023).

References

[Jansen and Spink2005] Bernard J. Jansen and Amanda Spink. 2005. An analysis of web searching by european alltheweb.com users. *Information Processing & Management*, 41(2):361–381.

[Kaplan et al.2021] Yohay Kaplan, Naama Krasne, Alex Shtoff, and Oren Somekh. 2021. Unbiased filtering of accidental clicks in Verizon Media native advertising. In *Proc. of the 30th ACM Int. Conf. on Information & Knowledge Management*, CIKM '21, page 3878–3887, New York, NY, USA. ACM.

[Scarone et al.2023] Bruno Scarone, Ricardo Baeza-Yates, and Erik Bernhardson. 2023. Understanding search behavior bias in wikipedia. *Fourth International Workshop on Algorithmic Bias in Search and Recommendation (Bias 2023) at the European Conference on Information Retrieval (ECIR)*, 4.

[Tolomei et al.2019] Gabriele Tolomei, Mounia Lalmas, Ayman Farahat, and Andrew Haines. 2019. You must have clicked on this ad by mistake! Data-driven identification of accidental clicks on mobile ads with applications to advertiser cost discounting and click-through rate prediction. *International Journal of Data Science and Analytics*, 7(1):53–66.

[Weinreich et al.2008] Harald Weinreich, Hartmut Obendorf, Eelco Herder, and Matthias Mayer. 2008. Not quite the average: An empirical study of web use. *ACM Trans. Web*, 2(1), Mar.

Notes

A. https://github.com/wikimedia/schemas-event-primary/blob/master/jsonschema/mediawiki/cirrussearch/request/0.0.1.yaml

B. https://wikitech.wikimedia.org/wiki/Analytics/Data_Lake/Content/Mediawiki_wikitext_current

C. https://wikitech.wikimedia.org/wiki/Analytics/Data_Lake/Traffic/Webrequest

D. https://github.com/wikimedia/analytics-refinery-source/blob/master/refinery-core/src/main/java/org/wikimedia/analytics/refinery/core/Webrequest.java#L277

E. https://wikitech.wikimedia.org/wiki/Analytics/Data_Lake/Traffic/Pageview_hourly

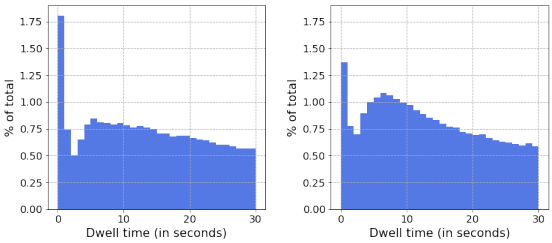
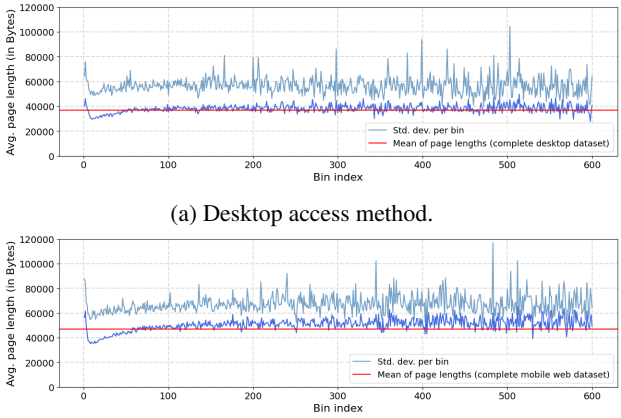


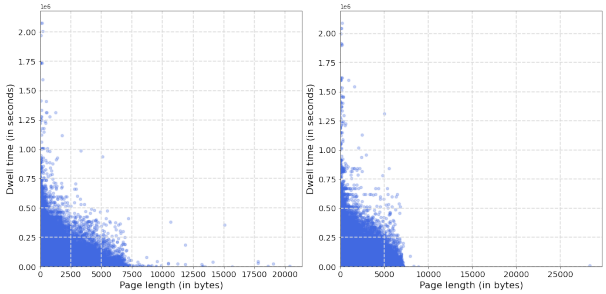
Figure 1: Dwell time histograms for 4-10/July/2022 (sub-range $[0s, 30s]$) using one second bins for desktop (left) and mobile web devices (right).



(a) Desktop access method.

(b) Mobile web access method.

Figure 2: Average page length per dwell time bin (sub-range $[0s, 600s]$).



(a) Desktop access method. (b) Mobile web access method.

Figure 3: Dwell time vs Page length - linear scale.

Result	Novel for Wikipedia users seg. by client type (abbr. c.t.)	Rediscovered to be valid in (website) search	Novel for web-site search
Accidental clicks phenomenon	✓ [Ind. of c.t.]	✓ [Prev.: mobile ads]	✓
Time a user spends on a Wikipedia page is ind. from page length	✓ [Ind. of c.t.]	✓ [Prev.: web search, we expand sample size]	✓
Ranking pos. clicked on is ind. from TTFC	✓ [Ind. of c.t.]	-	✓

Table 3: Summary of contributions.