PROBLEMS OF INFORMATION SOCIETY

Tiziano Piccardi, Robert West
School of Computer and Communication Sciences, EPFL
(École polytechnique fédérale de Lausanne), Lausanne, Switzerland

Miriam Redi
Wikimedia Foundation, France

Giovanni Colavizza
Laboratory of Digital Humanities, University of Amsterdam, Amsterdam, Netherlands

Quantifying Engagement with Citations on Wikipedia. (Part 2)

Abstract: Wikipedia is one of the most visited sites on the Web and a common source of information for many users. As an encyclopedia, Wikipedia was not conceived as a source of original information, but as a gateway to secondary sources: according to Wikipedia's guidelines, facts must be backed up by reliable sources that reflect the full spectrum of views on the topic. Although citations lie at the heart of Wikipedia, little is known about how users interact with them. To close this gap, we built client-side instrumentation for logging all interactions with links leading from English Wikipedia articles to cited references during one month, and conducted the first analysis of readers' interactions with citations. We find that overall engagement with citations is low: about one in 300 page views results in a reference click (0,29% overall; 0,56% on desktop; 0,13% on mobile). Matched observational studies of the factors associated with reference clicking reveal that clicks occur more frequently on shorter pages and on pages of lower quality, suggesting that references are consulted more commonly when Wikipedia itself does not contain the information sought by the user. Moreover, we observe that recent content, open access sources, and references about life events (births, deaths, marriages, etc.) are particularly popular. Taken together, our findings deepen our understanding of Wikipedia's role in a global information economy where reliability is ever less certain, and source attribution ever more vital.

3.5. General statistics of English Wikipedia

By the end of the data collection, English Wikipedia contained 5.8 M articles, 5.4 M (95%) of which were loaded at least once in our data sample, in a total of 7.4 M revisions. Out of these articles, 3.9 M (73%) contain at least one citation, linking to a total of 24 M distinct URLs.

Over the 4 weeks of data collection, we collected (at a 33% sampling rate) 1.5 B pageLoad events (62% from the mobile site and the rest from the desktop site). In Fig. 2a we report the (complementary cumulative) popularity distribution for the Wikipedia pages that were viewed at least once during the data collection period. The distribution is heavily skewed, with approximately 83% of the articles loaded fewer than 100 times in the 33% random sample (cf. Sec. 3.2), or fewer than 300 times when extrapolating to all data.

We observe a similar uneven distribution of page length (Fig. 2b), with the majority of articles being very short.

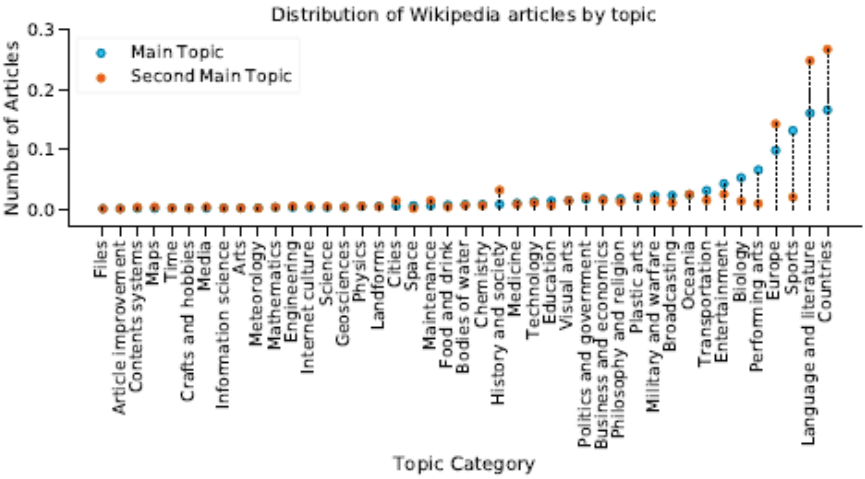

Figure 3. Distribution of most and second most prominent Wikipedia article topics (Sec. 3.5)

Fig. 2c shows that the distribution of article quality levels is also heavily skewed toward low quality levels: most articles are identified as “Stub” or “Start”, and fewer than 300 K articles are marked as “Good” or “Featured” articles.

Finally (Fig. 3), we find that a majority of articles are about geography or “Language and literature” (the latter including biographies), followed by topics related to sports and science.

4. RQ1: prevalence of citation interactions

After these preliminaries, we are now ready to address our first research question, which asks to what extent Wikipedia readers engage with citations.

4.1. Distribution of interaction types

We start by analyzing the relative frequency of the different citation events, as defined in Sec. 3.2. Over the month of data collection, we captured a total of 96 M citation events. Fig. 4 shows how these events distribute over the 5 event types, broken down by device type (mobile vs. desktop). We observe that most interactions with citations happen on desktop rather than mobile devices, despite the fact that the majority of page loads (62%) are made from mobile.

The interactions also distribute differently across types for mobile vs. desktop. The by far prevailing event on desktop is hovering over a footnote (fnHover) in order to display the reference text. Hovering requires a mouse, which is not available on most mobile devices, which in turn explains the low incidence of fnHover on mobile. In order to reveal the reference text behind a footnote, mobile users instead need to click on the footnote, which presumably explains why fnClick is the most common event on mobile.

Clicking external links outside of the References section at the bottom of the page (extClick) is the second most common event on both desktop and mobile, followed by clicks on citations from the References section (refClick). Finally, the upClick action, which lets users jump back from the References section to the locations where the citation is used in the main text, is almost never used.

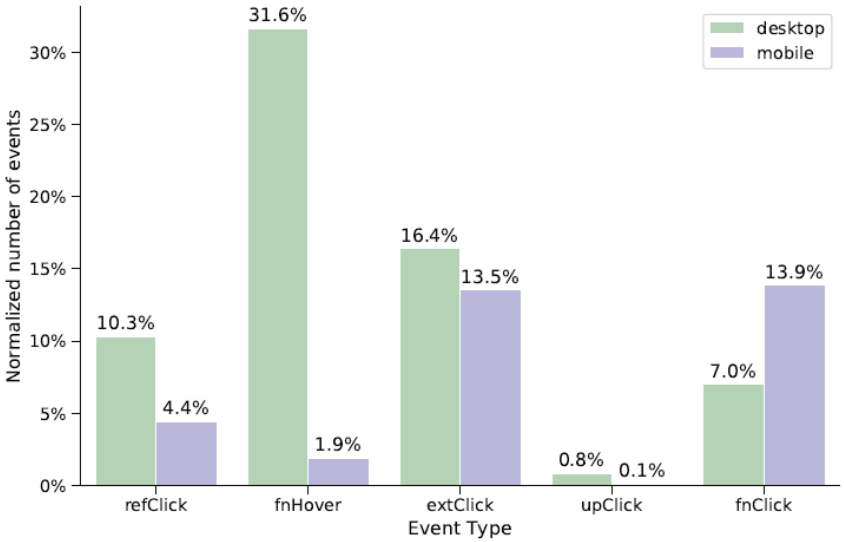

Figure 4. Relative frequency of citation-related events (Sec. 3.2), split into desktop (green, left bars) and mobile (blue, right bars) in April 2019 (Sec. 4.1)

4.2. Citation click-through rates

We now focus on the two prevalent interactions with citations, hovering over footnotes (fnHover) and leaving Wikipedia by clicking on citation links (refClick). (We do not dwell on extClick events, as they do not concern citations but other external links; cf. Sec. 3.2.)

First, we observe that, out of the 24 M distinct URLs that are cited across all articles in English Wikipedia, 93% of the URLs are never clicked during our month of data collection.

Next, we note that the global click-through rate (CTR) across all pages with at least one citation (gCTR, Eq. 1) is 0.29%; i.e., clicks on references happen on fewer than 1 in 300 page loads. Breaking the analysis up by device type, we observe again substantial differences between

desktop and mobile: on desktop the global CTR is 0.56%, over 4 times as high as on mobile, where it is only 0.13%.

The average page-specific CTR (pCTR, Eq. 3) is higher, at 1.1% for desktop and 0.52% for mobile. This is due to the fact that there are many rarely viewed pages (cf. Fig. 2a) with a noisy, high CTR.

After excluding pages with fewer than 100 page views, the global CTR is 0.67% on desktop, and 0.21% on mobile.

Engagement via footnote hovering is slightly higher, at a global footnote hover rate (gHR, Eq. 4) of 1.4%. The average page-specific footnote hover rate (pHR, Eq. 4) is 0.68% when including all pages with at least one clickable reference, and 1.1% when excluding pages with fewer than 100 page views*.

Given these numbers, we conclude that readers' engagement with citations is overall low.

4.3. Positional bias

Previous work has shown that users are more likely to click Wikipedia-internal links that appear at the top of a page [**42**]. To verify whether this also holds true for references, we sample one random page load with citation interactions per session and randomly sample one clicked and one unclicked reference for this page load. We then compute each reference's relative position in the page as the offset from the top of the page divided by the page length (in characters). Fig. 5, which shows the distribution of the relative position for clicked and unclicked references, reveals that users are more likely to click on references toward the top and (less extremely so) the bottom of the page.

4.4. Top clicked domains

Next, we investigate what are the most frequent domains at which users arrive upon clicking a citation.

Initially, we found that the most frequently clicked domain is archive.org (Internet Archive), with 882 K refClick events. Such URLs are usually snapshots of old Web pages archived by the Internet Archive's

---

* As mentioned in Sec. 4.1, hovering is not available on most mobile devices, so the hovering numbers pertain to desktop devices only.

Wayback Machine. To handle such cases, we extract the original source domains from wrapping archive.org URLs.

In Fig. 7 we report the top 15 domains by number of refClick events. The most clicked domain is google.com. Drilling deeper, we checked the main subdomains contributing to this statistic, finding that a significant proportion of clicks goes to books.google.com, which is providing partial access to printed sources. The second most clicked domain is doi.org, the domain for all scholarly articles, reports, and datasets recorded with a Digital Object Identifier (DOI), followed by (mostly liberal) newspapers (The New York Times, The Guardian, etc.) and broadcasting channels (BBC).

4.5. Markovian analysis of citation interactions

Whereas the above analyses involved individual events, we now begin to look at sessions: sequences of events that occurred in the same browser tab (as indicated by the session token; Sec. 3.2). Every session starts with a pageLoad event, and we append a special END event after the last actual event in each session.

By counting event transitions within sessions, we construct the first-order Markov chain that specifies the probability P(j | i) of observing event j right after event i, where i and j can take values from the event set introduced in Sec. 3.2 (pageLoad, refClick, extClick, fnClick, upClick, fnHover) plus the special END event.

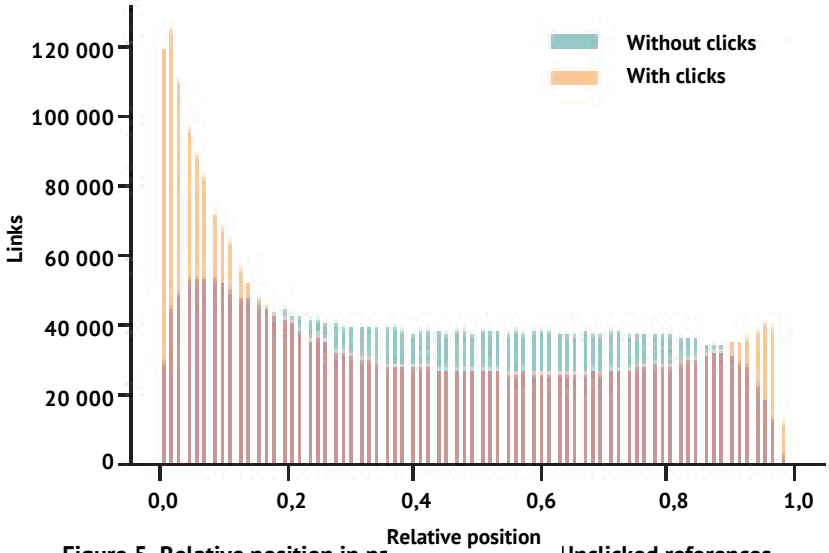

Figure 5. Relative position in page of clicked vs. Unclicked references, for references with hyperlinks (Sec. 4.3)

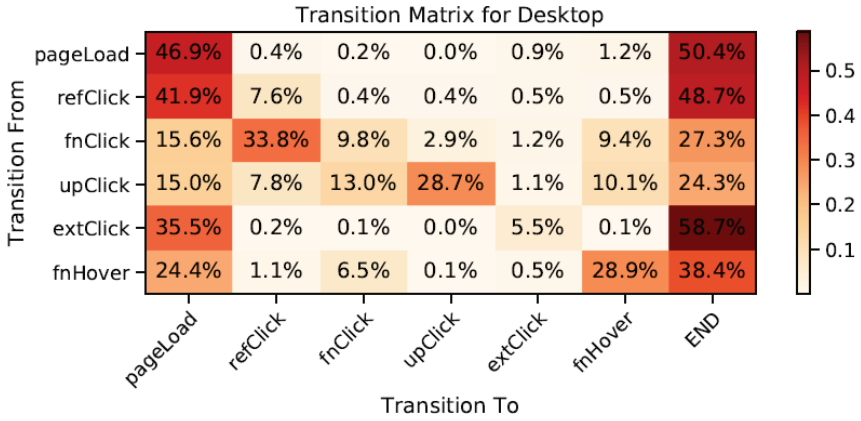

Figure 6a. Transition matrices of first-order Markov chains for desktop devices aggregating reader behavior with respect to citation events when navigating a Wikipedia article with references (Sec. 4.5)

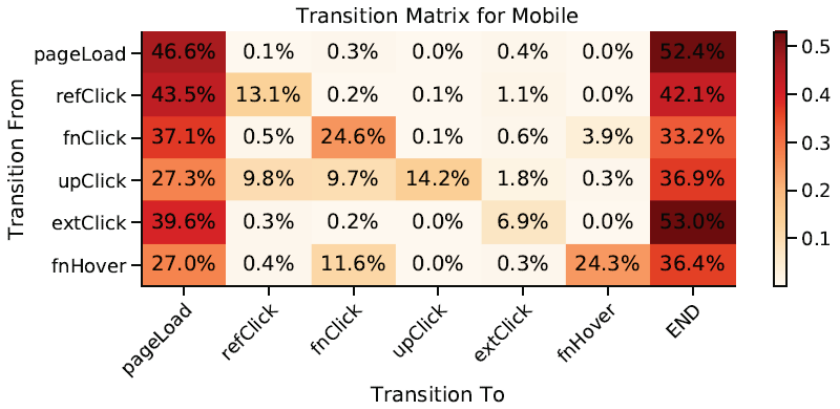

Figure 6b. Transition matrices of first-order Markov chains for mobile devices, aggregating reader behavior with respect to citation events when navigating a Wikipedia article with references (Sec. 4.5)

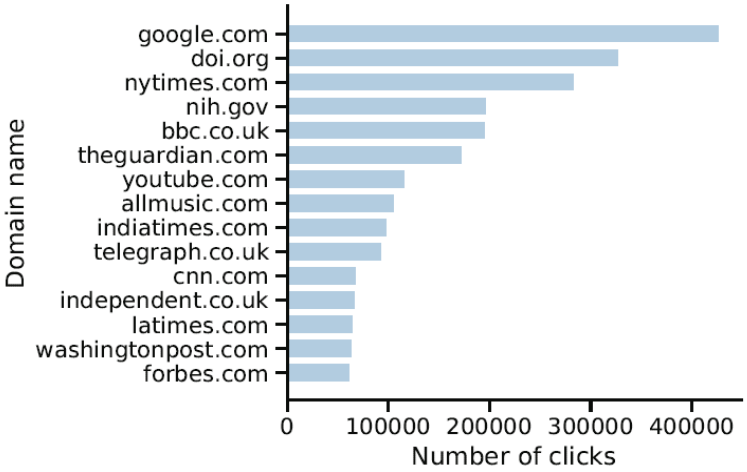

Figure 7. Top 15 domainnames appearing in English Wikipedia references (Sec. 4.4), sorted by number of clicks receivedduring April 2019

The transition probabilities are reported in Fig. 6. We observe that most reading sessions are made up of page views only: on both desktop and mobile, after loading a page, readers tend to end the session (with a probability of around 50%) or load another page in the same tab (47%). All citation-related events have a very low probability (at most 1.2%) of occurring right after loading a page.

On desktop, reference clicks become much more likely after footnote clicks (34%), and footnote clicks in turn become much more likely after footnote hovers (6.5%), hinting at a common 3-step motif (fnHover, fnClick, refClick), where the reader engages ever more deeply with the citation. Note, however, that this is not true for mobile devices, where, even after readers clicked on a footnote, the probability of also clicking on the citation stays low (0.5%).

Finally, reference clicks (refClick) are also common immediately after other reference clicks (8% on desktop, 13% on mobile). Note that for external links outside of the References section (extClick) we see a different picture: such external clicks are only rarely followed by interactions with citations (fnHover, fnClick, refClick), and in the majority of cases (59% on desktop, 53% onmobile) they conclude the session, suggesting that Wikipedia is in these cases commonly used as a gateway to external websites.

Information about the authors

Tiziano Piccardi – School of Computer and Communication Sciences, EPFL (École polytechnique fédérale de Lausanne), Lausanne, Switzerland

tiziano.piccardi@epfl.ch

Robert West – Assistant Professor, Data Science Laboratory, School of Computer and Communication Sciences, EPFL (École polytechnique fédérale de Lausanne), Lausanne, Switzerland

robert.west@epfl.ch

Miriam Redi – Research Scientist, Research Group, Wikimedia Foundation, France

miriam@wikimedia.org

Giovanni Colavizza – Assistant Professor, Laboratory of Digital Humanities, University of Amsterdam, Amsterdam, Netherlands

g.colavizza@uva.nl