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 3)

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.

5. RQ2: PAGE-LEVEL ANALYSIS OF CITATION INTERACTIONS

We now proceed to our second research question, which asks what features of a Wikipedia page predict whether readers will engage with the references it contains.

5.1. Predictors of reference clicks

As a first step, we perform a regression analysis. We train a logis-tic regression classifier for predicting whether a given pageLoad event will eventually be followed by a refClick event. To assem-ble the training set, we first find sessions with at least one (pos-itive) pageLoad followed by a refClick and at least one (negative) pageLoad not followed by a refClick, and make sure to include at most one such pair per session in order to avoid over-representing power users with extensive sessions. The dataset totals 938K pairs, which we split into 80% for training and 20% for testing.

As predictors we use the article's topic vector (with entries from [0, 1]; Sec. 3.4) and the quality label (Sec. 3.4), which we also normal-ize to a score in the range [0, 1] using the mapping from a previous study [**20**]. We did not use the number of references and the length of the page, as they are important features in the quality model and would cause collinearity issues due to their high correlation with quality (Pearson's correlation 0.81 and 0.75, respectively).

The resulting regression model has an area under the ROC curve (AUC) of 0.6 on the testing set. A summary of the 10 most predictive positive and negative coefficients is given in Fig. 8. By far the most important predictor — with a large negative weight — is the article's quality. Moreover, some topics are positive predictors (e.g., "Language and literature", which also includes all biographies, as well as "Internet culture"), while others are negative predictors (e.g., "Media", "Information science").

Given the importance of the quality feature in this first analysis, we now move to investigating its role in a more controlled study.

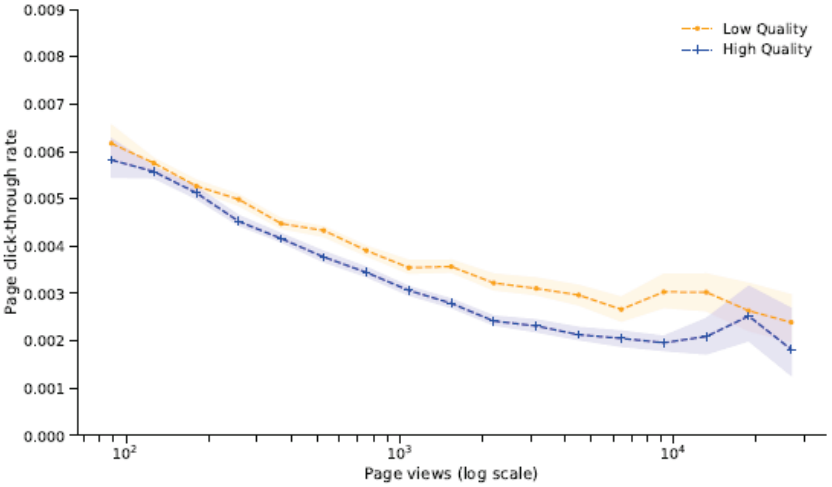

Figure 9. Comparison of page-specific click-through rate for low- (yellow) vs. high-quality (blue) articles, as function of popularity (Sec. 5.2). Error bands: bootstrapped 95% CIs.

5.2. Effects of page quality

To come closer to a causal understanding of the impact of an article's quality on readers' clicking citations in the article, we perform a matched observational study. The ideal goal would be to compare the page-specific CTR (Eq. 2) for pairs of articles — one of high, the other of low quality — that are identical in all other aspects.

Propensity score. Finding such exact matches is unrealistic in practice, so we resort to propensity score matching [**4**], which provides a viable solution. The propensity score specifies the proba-bility of being treated as a function of the observed (pre-treatment) covariates. Crucially, data points with equal propensity scores have the same distribution over the observed covariates, so matching treated to untreated points based on propensity scores will balance the distribution of observed covariates across treatment groups.

In our setting, we define being of high quality as the treatment and estimate propensity scores via a logistic regression that uses topics, length, number of citations, and popularity as observed co-variates in order to predict quality as the binary treatment variable. We consider as low-quality all articles tagged as Stub or Start (74% of the total; Fig. 2c), and as high-quality the rest. Articles without a refClick or fewer than 100 pageLoad events are discarded in order to avoid noisy estimates of the page-specific CTR. This leaves us with 854K articles.

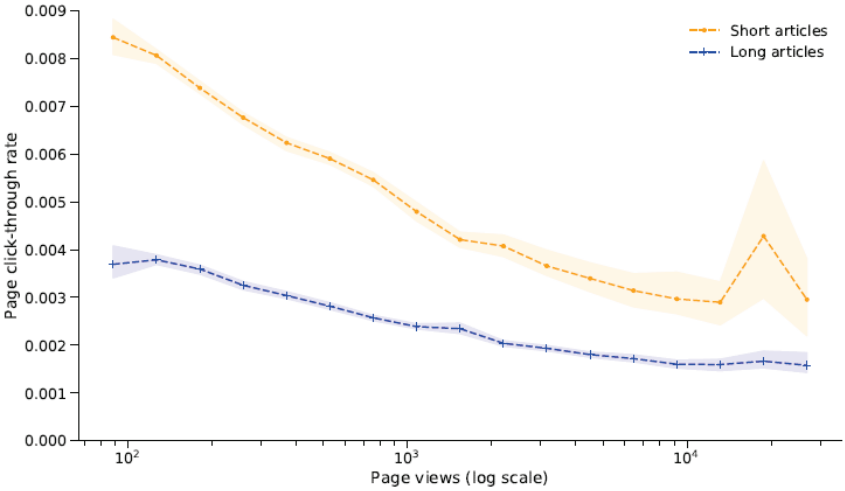

Figure 10. Comparison of page-specific click-through rate for short (yellow) vs. long (blue) articles, as function of pop-ularity (Sec. 5.3). Error bands: bootstrapped 95% CIs.

Matching. We compute a matching (comprising 198K pairs) that minimizes the total absolute difference of within-pair propensity scores, under the constraint that the length of matched pages should not differ by more than 10%. This constraint is necessary to ascertain balance

on the page length feature because page length is so highly correlated with quality (Pearson correlation 0.81; cf. Sec. 5.1). After matching, we manually verify that all observed covariates, including page length, are balanced across groups.

Results. Fig. 9 visualizes the average page-specific CTR for articles of low (yellow) and high (blue) quality as a function of article popularity. We can observe that the CTR of low-quality articles significantly surpasses that of high-quality articles across all levels of popularity. In interpreting this result, it is important to recall that page length is one of the most important features in ORES [**20**], the quality-scoring model we use here. As we control for page length, the gap observed in Fig. 9 may be attributed to the remaining features used by ORES, such as the presence of an infobox, the number of images, and the number of sections and subsections.

We hence dedicate our next, final page-level analysis to estimat-ing the impact of page length alone on page-specific CTR.

5.3. Effects of page length

In order to measure the effect of page length on CTR, we take a two-pronged approach, first via a cross-sectional study using propensity scores, and second via a longitudinal study.

Cross-sectional study. First, we conduct a matched study based on propensity scores analogous to Sec. 5.2, but now with page length as the treatment variable (using the longest and the shortest 40% of articles as treatment groups), and all other features (except quality) as observed covariates. Matching yields 683K pairs, and we again manually verify covariate balance across treatment groups.

The average page-specific CTR of short articles (0.68%) is more than double that of long articles (0.27%; $p \ll 0.001$ in a two-tailed Mann–Whitney U test). Moreover, as seen in Fig. 10, this relative difference obtains across all levels of article popularity.

Longitudinal study. While in the above cross-sectional study propensity score matching ensures that the covariates of long vs. short articles are indistinguishable at the aggregate treatment group level, it does not necessarily do so at the pair level. Also, we did not include as observed covariates features describing the users who read the respective articles, and it might indeed be the case that users with a liking for short, niche articles also have a higher probability of clicking citations. In order to mitigate the danger of such remaining potential confounds and achieve even finer control, we now conduct a longitudinal study to assess how a variation in length of the same article impacts its CTR.

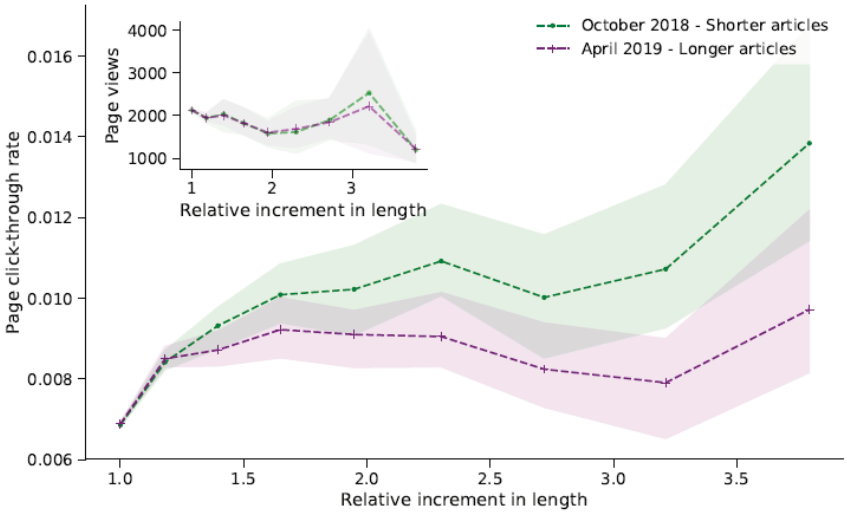

Figure 11. Comparison of page-specific click-through rate of shorter (green) vs. longer (purple) revisions of identical arti-cles, as function of length ratio (Sec. 5.3). Inset: popularity as function of length ratio. Error bands: bootstrapped 95% CIs.

To do so, we select all articles that grew in length between Octo-ber 2018 and April 2019, our two data collection periods (Sec. 3.2). To control for the effect of page popularity, which was observed to negatively correlate with CTR (Fig. 9 and 10), we assign a popular-ity level to each article by binning page view counts into deciles and discard articles whose popularity level has changed between the two periods. This way, we obtain a set of 120K articles with matched long and short revisions.

By grouping these articles by the length ratio of their two revi-sions and plotting this ratio against the CTR for the long (purple) vs. short (green) versions (Fig. 11), we provide a further strong indi-cator that page length causally decreases the prevalence of citation clicking. According to a Mann–Whitney U test, the CTR difference between long and short revisions is statistically significant with $p < 0.05$ starting from a length increase of 17%, and with $p < 0.01$ from 31%. In addition, to verify that the effect is not confounded by a concomitant change in article popularity, the inset plot in Fig. 11 shows that the popularity indeed stays constant between revisions.

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