ПРОБЛЕМЫ ИНФОРМАЦИОННОГО ОБЩЕСТВА

УДК 001.811
DOI: 10.33186/1027-3689-2020-11-147-156

Тициано Пикарди, Роберт Вест
Школа компьютерных и коммуникационных наук
Федеральной политехнической школы Лозанны, Швейцария

Мириам Реди
Фонд Викимедия, Франция

Джованни Колавицца
Лаборатория цифровых общественных наук
Университета Амстердама, Нидерланды

Количественные характеристики работы с цитатами в Википедии. (Часть 3)

Аннотация: Википедия является одним из самых посещаемых сайтов в интернете и распространённым источником информации для многих пользователей. В качестве энциклопедии Википедия задумывалась не как источник оригинальной (окончательной) научной информации, а, скорее, как ворота к более глубоким и точным источникам. В соответствии с базовыми принципами Википедии факты должны быть подкреплены надёжными источниками, которые отражают полный спектр всех мнений по данной теме. Хотя цитаты лежат в основе функционирования Википедии, пока мало что известно о том, как пользователи работают с ними. Чтобы закрыть этот пробел, мы создали клиентские (пользовательские) инструменты для ведения записей (журналов) всех взаимодействий со ссылками, идущими из англоязычных статей Википедии на цитируемые ссылки в течение одного месяца, и провели первый анализ взаимодействия читателей с цитатами.

Результаты показывают, что в целом вовлечённость в цитаты низкая. Около 300 просмотров страниц приводят к входу на одну ссылку – это составляет всего 0,29%; в том числе 0,56% при работе с настольным компьютером (на рабочем столе) и 0,13% при работе на мобильных устройствах. Сопоставление факторов, связанных с переходами по ссылке, показывает, что переходы происходят чаще на более коротких страницах и на страницах относительно низкого качества. Исходя из этого можно предположить, что ссылки чаще всего требуются, когда

Википедия не содержит информацию, которую ищет пользователь. Кроме того, мы обратили внимание, что источники открытого доступа и ссылки о жизненных событиях (рождения, смерти, браки и т.д.) особенно популярны.

Собранные воедино, наши выводы углубляют понимание роли Википедии в глобальной информационной экономике, где надёжность становится всё менее определённой, а значение источников становится всё более важным.

Справочный формат ACM для ссылок:

Тициано Пикарди, Мириам Реди, Джованни Колавицца и Роберт Вест. 2020.

Количественная оценка взаимодействия с цитатами в Википедии. В трудах: Веб-конференция 2020 (WWW'20), 20–24 апреля 2020 г., Тайбэй, Тайвань. ACM, Нью-Йорк, штат Нью-Йорк, США. 12 с. https://doi.org/10.1145/3366423.3380300.

Ключевые слова: цитирование, гиперссылки, примечания, справки, Википедия, математическая статистика, поведение пользователей.

6. Анализ работы с цитатами на уровне страницы

Переходим ко второму вопросу нашего исследования: какие особенности страницы Википедии показывают, будут ли читатели работать со ссылками, которые эта страница содержит?

Предикторы[1] ссылочных кликов

В качестве первого шага мы выполняем регрессионный анализ. Настраиваем логистический классификатор регрессии, чтобы спрогнозировать, является ли данный *pageLoad* таким, за которым в конечном итоге последует событие *refClick.* Для формирования тренировочного набора мы сначала отбираем сеансы с хотя бы одним (положительным) событием *pageLoad,* за которым последовало событие *refClick,* и хотя бы одно событие (отрицательное) *pageLoad,* которое не сопровождается *refClick,* и обязательно включаем только одну такую пару за сеанс, чтобы избежать чрезмерного влияния продвинутых пользователей с обширными сессиями. Полученный таким образом набор данных насчитывает 938 тыс. пар, который мы разделили следующим образом: 80% для обучения и 20% для тестирования.

[1] Предиктор – прогностический параметр; средство прогнозирования, экстраполяционная функция. – *Примеч. пер.*

В качестве предикторов используем вектор темы статьи (с записями из диапазона [0, 1]) и знак качества, который мы также нормируем в диапазоне [0, 1] с помощью отображения из предыдущего исследования [**20**]. Мы не использовали такие параметры, как количество ссылок и длина страницы, так как они являются важными характеристиками в качественной модели и вызовут проблемы коллинеарности из-за их высокой корреляции с категорией качества (соотношение Пирсона 0,81 и 0,75 соответственно). Полученная регрессионная модель имеет площадь под кривой *ROC*[2] *(AUC)* 0,6 на испытательном комплекте.

Итоговые 10 самых значимых предсказательных положительных и отрицательных факторов приведены на рис. 8.

Наиболее важный предиктор (с большим отрицательным весом) – качество статьи. Более того, некоторые темы являются положительными предикторами (например, «Язык и литература», которая также включает в себя все биографии, а также «интернет-культура»), в то время как другие – негативными предикторами (например, «Медиа», «Информатика»). Учитывая важность фактора качества в этом первом анализе, мы переходим к изучению его роли в более детальном исследовании.

Влияние качества страницы

Чтобы лучше понять влияние качества статьи на количество кликов читателей в статье, нужно выполнить сравнительное и сопоставимое обсервационное исследование. Идеально было бы сравнивать показатель *CTR* конкретной страницы (уравнение 2) для пары статей: одна – высокого качества и другая – низкого качества, которые идентичны во всех других аспектах.

---

[2] *ROC*-анализ тесно связан с бинарной логистической регрессией и применяется для оценки качества моделей: позволяет аналитику выбрать модель с наилучшей прогностической силой. Площадь под *ROC*-кривой *AUC (Area Under Curve)* является агрегированной характеристикой качества классификации, не зависящей от соотношения цен ошибок. Чем больше значение *AUC,* тем «лучше» модель классификации. Коэффициент корреляции Пирсона характеризует существование линейной зависимости между двумя величинами. Его называют также теснотой линейной связи.

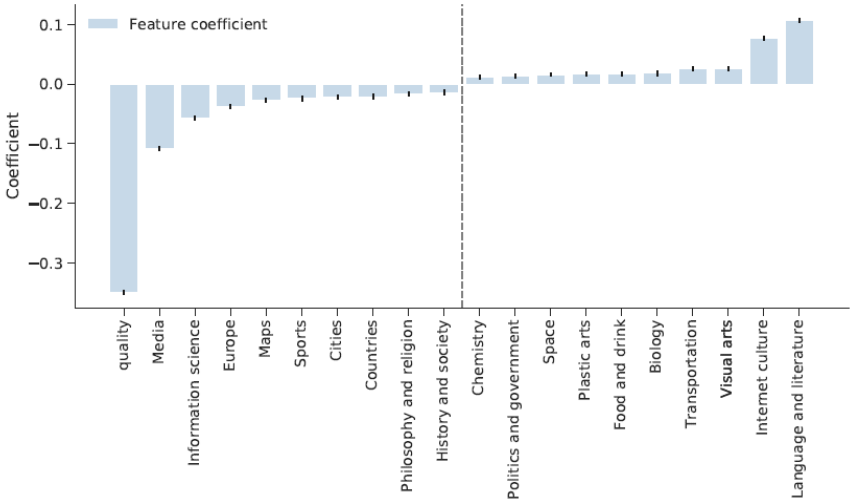

Рис. 8. Вклад различных факторов в регрессионную модель, оценивающую вероятность того, что произойдёт после загрузки страницы. (Показаны 10 позитивных факторов: слева направо – качество, среда, информатика, Европа, карты, спорт, города, страны, философия и религия, история и общество; и 10 негативных факторов: слева направо – химия, политика и управление, космос, пластические искусства, пища и напитки, биология, перевозки, визуальные искусства, культура интернета, язык и литература)

Оценка склонности. Найти такие точные совпадения на практике нереально, поэтому мы прибегаем к сопоставлению баллов склонности [**4**], которое обеспечивает достаточно практичное жизнеспособное решение. Оценка склонности определяет вероятность результатов обработки каких-либо параметров как функция наблюдаемых ковариат[3], иначе говоря, по результатам наблюдения за какими-то известными

[3] Ковариа́ция (корреляционный момент, ковариационный момент) – в теории вероятностей и математической статистике мера линейной зависимости двух случайных величин. Ковариата – это переменная, которая может влиять на взаимосвязь между изучаемыми переменными.

параметрами мы можем сделать вывод, что и ковариаты поведут себя аналогичным образом. Важно отметить, что данные с равными показателями склонности имеют одинаковое распределение по наблюдаемым ковариатам, поэтому соответствие обработанных и необработанных точек на основе баллов склонности будет сбалансированным распределением наблюдаемых ковариат по группам обработки.

В нашей ситуации мы определяем высокое качество как обработку и оцениваем параметры склонности с помощью логистической регрессии, которая использует такие категории, как тематика, длина, количество цитирований и популярность как наблюдаемые ковариаты, чтобы предсказать качество как бинарную переменную обработки. Мы считаем некачественными все статьи, помеченные как *Stub* («Отбросы, затычка») или *Start* («Начальный уровень») (74% от общего количества, рис. 2в), а качественными – всё остальное. Статьи без *refClick* или менее 100 событий *pageLoad* отбрасываются по порядку, чтобы избежать ложных оценок *CTR* для конкретной страницы. Таким образом, в нашем распоряжении остаётся 854 тыс. статей.

Сопоставление (Matching)[4]. Мы вычисляем соответствие (в массиве, состоящем из 198 тыс. пар), которое минимизирует общую абсолютную разницу склонностей внутри пары, при условии, что длина сопоставляемых страниц не должна отличаться более чем на 10%. Это ограничение необходимо для сохранения баланса по категории длина страницы, потому что длина страницы очень сильно коррелирует с качеством (корреляция Пирсона 0,81). Затем, сопоставляя, мы вручную проверяем, что все наблюдаемые ковариаты, в том числе длина страницы, сбалансированы по группам.

---

[4] *Propensity Score Matching PSM* – метод корректировки исходных данных для получения более достоверных результатов сравнения групп наблюдений статистическими методами. Дословный перевод названия метода – «сопоставление оценок склонностей», также в литературе встречаются следующие варианты перевода: «метод отбора подобного по коэффициенту склонности», «отбор подобного по вероятности (склонности)», «псевдорандомизация», «метод подбора по индексу соответствия». Впервые метод *PSM* был представлен в публикации 1983 г., активное использование получил в последнее десятилетие.

Полученные результаты. На рис. 9 показан средний *CTR* для статей низкого качества (жёлтая кривая) и высокого качества (голубая кривая) как функция популярности статьи. Мы можем наблюдать, что *CTR* некачественных статей значительно превосходит *CTR* высококачественных на всех уровнях популярности. При интерпретации этого результата важно помнить, что длина страницы является одной из наиболее важных функций в *ORES* [**20**], той модели оценки качества, которую мы используем здесь. Поскольку мы собираемся изучать фактор длины страницы, наблюдаемый на рис. 9 разброс можно отнести к остальным параметрам, учитываемым в *ORES,* таким как наличие инфобокса, количество изображений, а также количество разделов и подразделов. Поэтому мы посвящаем наш следующий, завершающий, анализ на уровне страницы оценке влияния только длины страницы на *CTR* для конкретной страницы.

Влияние длины страницы

Для того, чтобы оценить влияние длины страницы на *CTR,* мы применяем два метода: перекрёстное исследование с использованием параметров склонности, а затем – продольное исследование.

Перекрёстное исследование. Во-первых, мы проводим совместное исследование на основе показателей склонности, но теперь длина страницы является переменной для обработки (в качестве групп обработки используются самые длинные 40% статей и самые короткие 40% статей), а все остальные параметры (кроме качества) воспринимаются как наблюдаемые ковариаты. Для сопоставления сформировано 683 тыс. пар, и мы снова вручную проверяем ковариатный баланс по группам обработки. Средний *CTR* для коротких статей для конкретной страницы – 0,68%, т.е. вдвое больше, чем у длинных статей (0,27%; $p \ll 0{,}001$ в двустороннем *U*-тесте Манна – Уитни[5]).

Более того, как видно на рис. 10, эта относительная разница достигается на всех уровнях популярности статьи.

---

[5] *U*-критерий Манна – Уитни (*Mann – Whitney U-test*) – статистический критерий, используемый для оценки различий между двумя независимыми выборками по уровню какого-либо признака, измеренного количественно. Позволяет выявлять различия в значении параметра между малыми выборками.

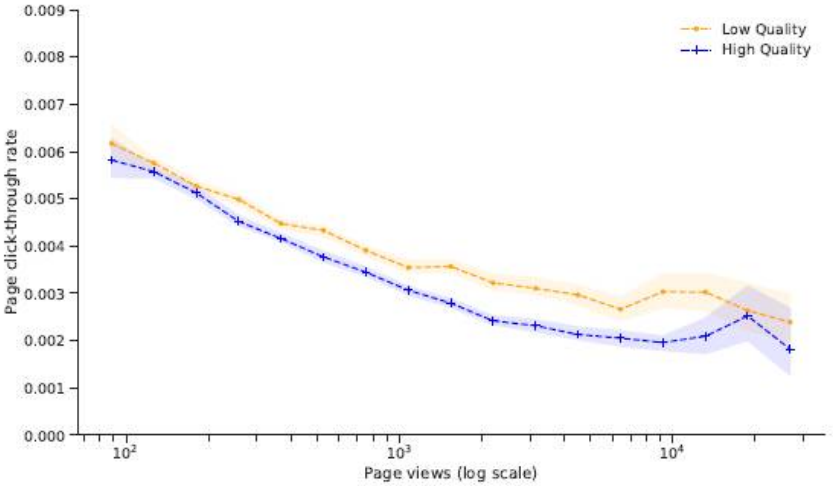

Рис. 9. Сравнение темпа переходов для низкокачественных (жёлтая кривая) и высококачественных (голубая кривая) статей одного и того же типа (горизонтальная ось – количество просмотров статей в логарифмическом масштабе, вертикальная ось – темп переходов; полоса ошибок соответствует доверительному интервалу[6] 95%)

[6] В статистике доверительный интервал – особая форма оценки определённого параметра. При использовании этого метода вместо одного значения задаётся весь интервал допустимых значений параметра, а также допускается вероятность того, что действительное (неизвестное) значение параметра будет находиться в этом интервале. Доверительный интервал основан на наблюдениях от выборки и, следовательно, отличается от выборки к выборке. Вероятность того, что параметр будет в интервале, называется уровнем достоверности. Очень часто этот показатель представлен в процентах. Доверительный интервал всегда указывается вместе с уровнем достоверности (https://en.wikipedia.org/wiki/Confidence_interval).

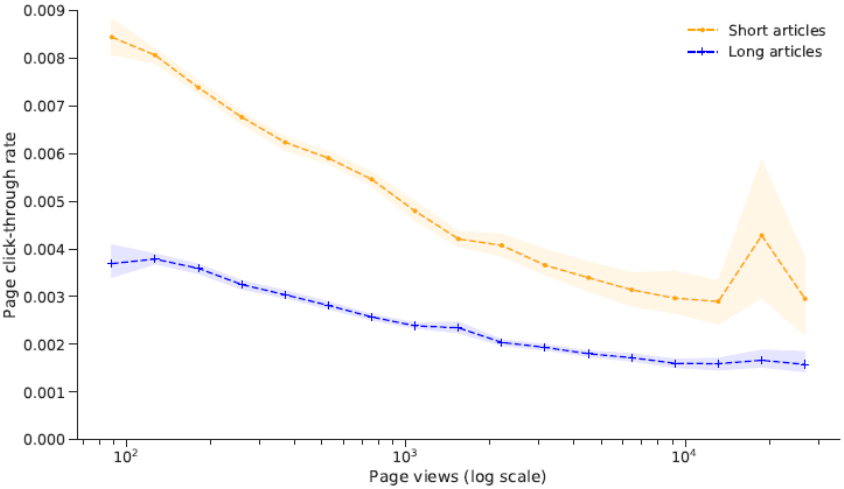

Рис. 10. Сравнение темпа переходов для коротких (жёлтая кривая) и длинных (голубая кривая) статей как функция популярности (горизонтальная ось – количество просмотров статей в логарифмическом масштабе, вертикальная ось – темп переходов кликов; полоса ошибок соответствует доверительному интервалу 95%)

Продольное исследование[7]. Хотя в приведённом выше перекрёстном исследовании сопоставление оценок склонностей гарантирует, что ковариаты длинных и коротких статей неразличимы на уровне групповой обработки, это не всегда верно на уровне пары. Кроме того, мы не включили в рассмотрение наблюдаемых ковариатических функций те, что описывают самих пользователей, – тех, кто читает соответствующие статьи. И это действительно может быть важно: если пользователям нравятся короткие, нишевые статьи, есть высокая вероятность того, что они нажмут на цитаты. Для того чтобы смягчить потенциальную опасность такой помехи и достичь ещё более точного результата, мы проводим продольное исследование с целью оценить, как изменение длины одной и той же статьи влияет на её рейтинг кликов.

[7] Продольное исследование (или продольное обследование, или групповое исследование) включает повторные наблюдения одних и тех же переменных в течение коротких или длительных периодов времени, т.е. используются продольные данные.

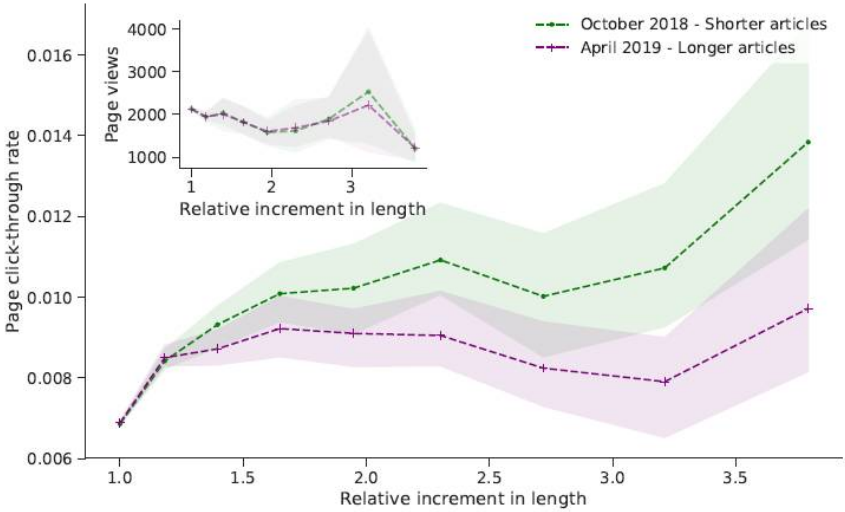

Рис. 11. Сравнение темпа переходов для коротких (зелёная кривая) и длинных (розовая кривая) статей (горизонтальная ось – относительное увеличение длины статьи, вертикальная ось – темп переходов). На вставке – популярность (востребованность) как функция длины статьи (горизонтальная ось – относительное увеличение длины статьи, вертикальная ось – число просмотров; полоса ошибок соответствует доверительному интервалу 95%)

Для этого мы отбираем все статьи, которые увеличились в длине в период с октября 2018 г. по апрель 2019 г. – за два периода сбора данных. Чтобы компенсировать влияние популярности страницы, которая отрицательно влияет на *CTR* (рис. 9 и 10), мы группируем уровни популярности статьи путём разбивки количества просмотров страниц на группы в децилях и отбрасываем статьи, уровень популярности которых изменился между двумя периодами. Таким образом, получаем набор из 120 тыс. статей с соответствующими длинными и короткими версиями.

Сгруппировав эти статьи по соотношению длины их двух периодов наблюдений и построив *CTR* для длинных (розовая кривая) по сравнению с короткими (зелёная кривая) версиями (рис. 11), мы получаем ещё один сильный индикатор того, что длина страницы заметно снижает клики цитирования. Согласно *U*-критерию Манна – Уитни, разница *CTR* между длинными и короткими ревизиями статистически значима с $p < 0,05$, начиная с увеличения длины на 17%, и с $p < 0,01$ с увеличения длины на 31%. Кроме того, чтобы убедиться, что эффект не смешивается с сопутствующим изменением популярности статьи, картинка на вставке на рис. 11 показывает: популярность действительно остаётся постоянной между пересмотрами.

Список литературы (70 позиций) представлен по адресу https://doi.org/10.1145/3366423.3380300.

(Продолжение в «НТБ» № 1 2021 г.)

Перевод А. И. Земскова, ГПНТБ России

Информация об авторах

Тициано Пикарди – Школа компьютерных и коммуникационных наук Федеральной политехнической школы Лозанны, Швейцария

tiziano.piccardi@epfl.ch

Роберт Вест – доцент лаборатории научных данных Школы компьютерных и коммуникационных наук Федеральной политехнической школы Лозанны, Швейцария

robert.west@epfl.ch

Мириам Реди – исследователь в научной группе Фонда Викимедия, Франция

miriam@wikimedia.org

Джованни Колавицца – доцент Лаборатории цифровых общественных наук – Университета Амстердама, Нидерланды

g.colavizza@uva.nl

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.

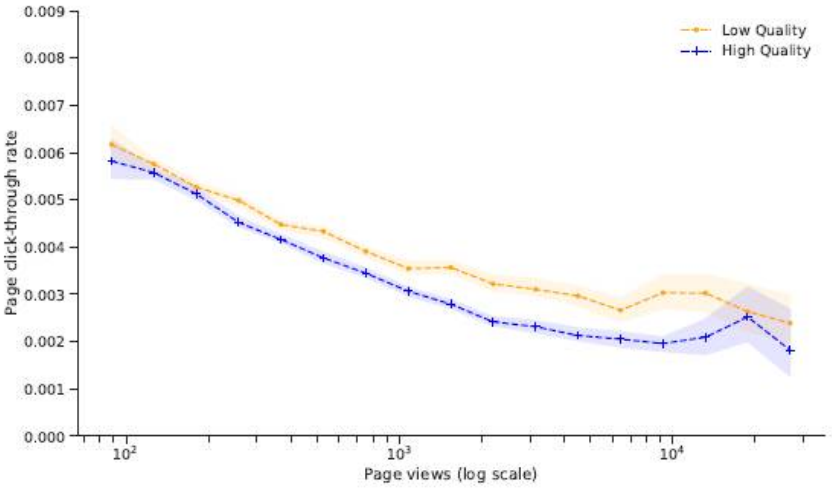

Figure 9. Comparison of page-specific click-through rate for low- (yellow) vs. high-quality (blue) articles, as function of popularity (Sec. 5.2). Error bands: bootstrapped 95% CIs.

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.

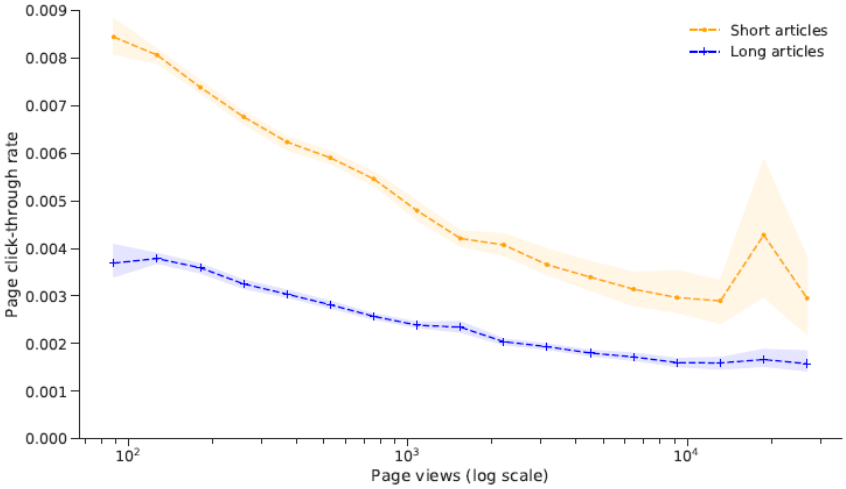

Figure 10. Comparison of page-specific click-through rate for short (yellow) vs. long (blue) articles, as function of pop-ularity (Sec. 5.3). Error bands: bootstrapped 95% CIs.

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.

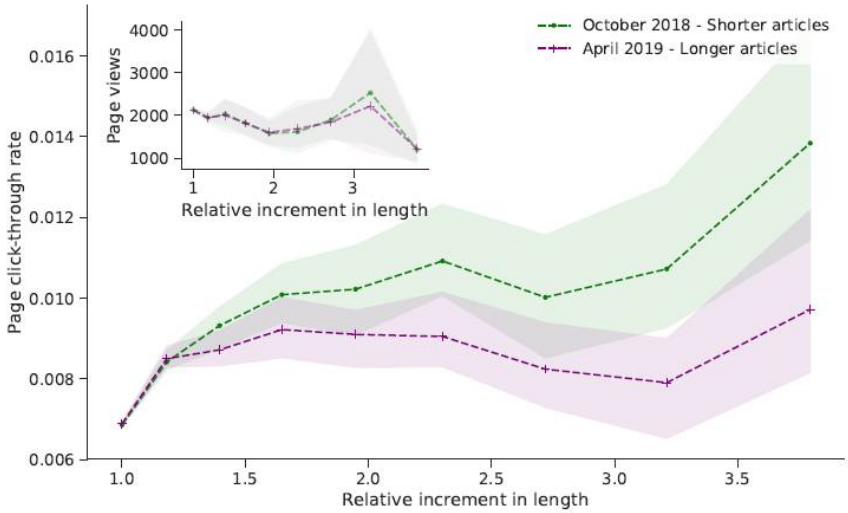

Figure 11. Comparison of page-specific click-through rate of shorter (green) vs. longer (purple) revisions of identical arti-cles, as function of length ratio (Sec. 5.3). Inset: popularity as function of length ratio. Error bands: bootstrapped 95% CIs.

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