

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 АСМ 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. АСМ, 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?

Предикторы¹ 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.

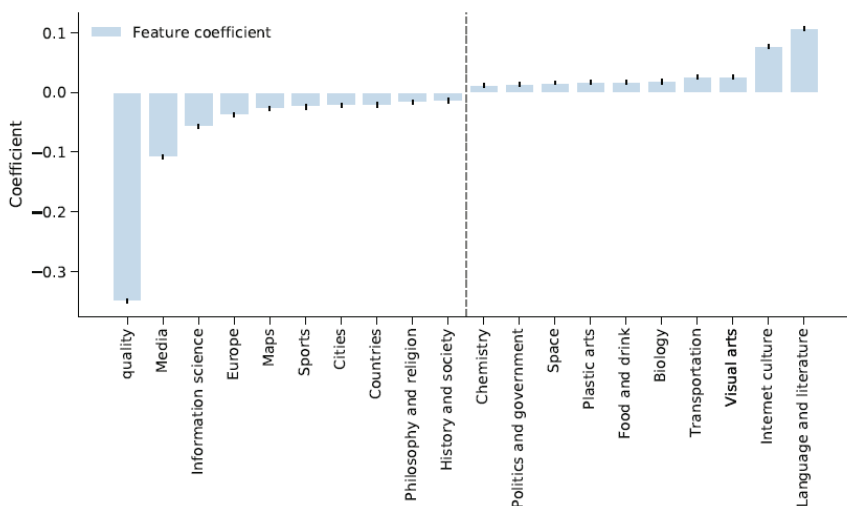


Рис. 8. Вклад различных факторов в регрессионную модель, оценивающую вероятность того, что произойдёт после загрузки страницы.
(Показаны 10 позитивных факторов: слева направо – качество, среда, информатика, Европа, карты, спорт, города, страны, философия и религия, история и общество; и 10 негативных 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 ковариат³, иначе говоря, по результатам наблюдения за какими-то известными

³ Ковариация (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 – Уитни⁵).

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.

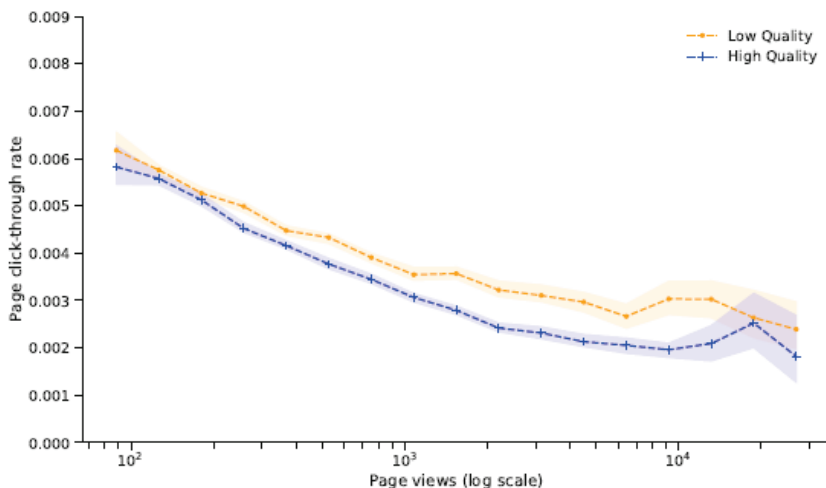


Рис. 9. Сравнение темпа переходов для низкокачественных (жёлтая кривая) и высококачественных (голубая кривая) статей одного и того же типа (горизонтальная ось – количество просмотров статей в логарифмическом масштабе, вертикальная ось – темп переходов; полоса ошибок соответствует доверительному интервалу⁶ 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).

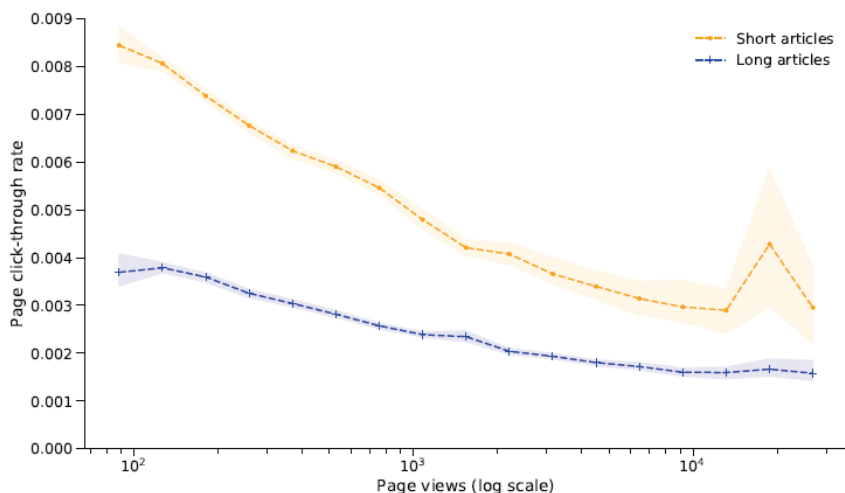


Рис. 10. Сравнение темпа переходов для коротких (жёлтая кривая) и длинных (голубая кривая) статей как функция популярности (горизонтальная ось – количество просмотров статей в логарифмическом масштабе, вертикальная ось – темп переходов кликов; полоса ошибок соответствует доверительному интервалу 95%)

*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.

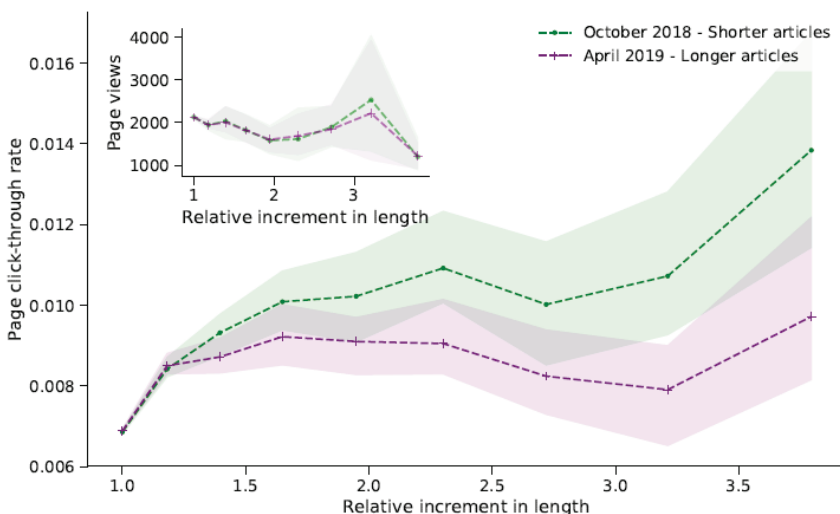


Рис. 11. Сравнение темпа переходов для коротких (зелёная кривая) и длинных (розовая кривая) статей (горизонтальная ось – относительное увеличение длины статьи, вертикальная ось – темп переходов). На вставке – популярность (востребованность) как функция длины статьи (горизонтальная ось – относительное 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 Колавица – доцент Лаборатории цифровых общественных наук – Университета Амстердама, Нидерланды

g.colavizza@uva.nl