УДК 001.811
DOI: 10.33186/1027-3689-2020-9-95-108

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 1)

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 Пиккарди, Мириам Реди, Джованни Колавицца и Роберт Вест. 2020.

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

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

Это коллективный труд группы молодых специалистов, имеющих отношение к Школе компьютерных и коммуникационных наук Федеральной политехнической школы Лозанны (*The School of Computer and Communication Sciences of the École polytechnique fédérale de Lausanne – EPFL*), подготовленный ими как доклад на конференции в Тайбэе (Тайвань) в апреле 2020 г. и размещённый в системе *Arxive* Корнельского университета, США, под лицензией *Creative Commons Attribution 4.0 International (CC-BY 4.0)*. Авторы оставили за собой право распространять работу на своих персональных и корпоративных сайтах с соответствующей атрибуцией. WWW'20, 20–24 апр. 2020 г., Тайбэй, Тайвань © 2020 IW3C2 (Международный комитет по Всемирной паутине), опубликовано под лицензией *Creative Commons CC-BY 4.0*. ACM ISBN 978-1-4503-7023-3/20/04. https://doi.org/10.1145/3366423.3380300.

Вряд ли можно назвать то, что изучали эти специалисты, цитированием. Скорее, речь идёт о многоступенчатом уточнении информации и предмета, интересующего читателя. Это происходит за счёт многочисленных ссылок и разъяснений, к которым может при желании обратиться пользователь. Изобилие этих добавок поражает. Пользователю предоставляется возможность многофасетной детализации интересующей его проблемы. Поведение пользователей Википедии в ходе процесса постепенного, последовательного познания – вот что изучалось в этой работе.

При переводе мы посчитали необходимым вставить несколько примечаний, связанных с терминами из математической статистики, – комплементарная интегральная функция распределения, Марковский анализ, предиктор, площадь под *ROC*-кривой *AUC,* коэффициент корреляции Пирсона, ковариация, доверительный интервал, метод корректировки исходных данных, *U*-критерий Манна – Уитни. Мы их выделили курсивом.

1. Введение

Википедия – самая большая энциклопедия из когда-либо созданных – является плодом совместных усилий большого редакторского коллектива, самоуправляется посредством согласованной политики на основе руководящих принципов [**7**, **16**]. Благодаря упорной работе сообщества редакторов содержание Википедии, как правило, качественное и актуальное [**25**, **45**] и заслуживает доверия как источник нейтральной, непредвзятой информации [**35**].

Встроенные ссылки (цитаты) Википедии являются ключевым механизмом для мониторинга и поддержания высокого качества. Мы используем термины *справка*, *ссылка* (*reference*) и *цитирование* (*citation*) как взаимозаменяемые. Базовая политика Википедии в отношении содержания требует, чтобы «люди, использующие энциклопедию, могли убедиться, что информация поступает из надёжного источника»

(https://en.wikipedia.org/wiki/Wikipedia:Verifiability, https://en.wikipedia.org/wiki/Wikipedia:Reliable_sources) и цитаты являются основным способом связать утверждение с его источниками. Отличительная черта Википедии – «действенность» многих цитат: они часто снабжены гиперссылками на цитируемый материал, доступный в интернете.

В результате роль Википедии в интернете была определена как «мост к следующему уровню академических ресурсов» [**19**] и «ворота, через которые миллионы людей теперь ищут доступ к знанию» [**11**]. И тем не менее остаётся открытым вопрос: в какой мере читатели Википедии действительно пересекают мост к знаниям и получают доступ к более широким источникам знания, упомянутым в энциклопедии? Учитывая коллективную и открытую природу Википедии, возможность количественно оценить взаимодействие пользователей с контентом и его вспомогательными источниками имеет решающее значение для постоянного улучшения энциклопедии и её роли в формировании самокритичного общества.

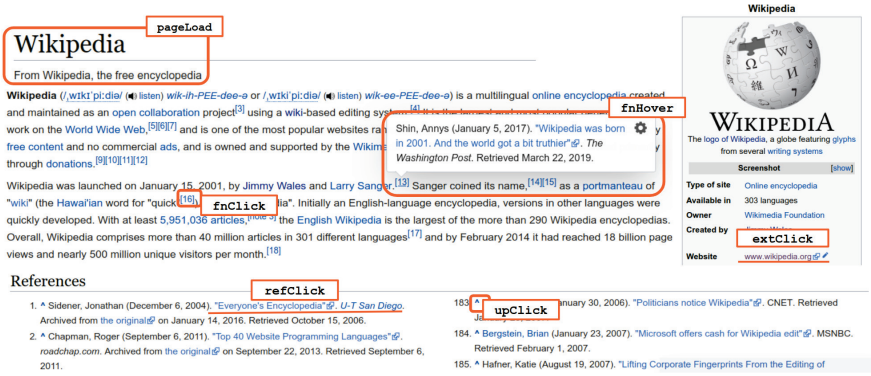
Понимание способа взаимодействия читателей с цитатами позволяет лучше оценить роль редакторов и политики Википедии в поддержании высокого качества информации, измерить общественный спрос на вторичные источники и дать потенциальные рекомендации для повышения интереса общественности к ссылкам. Данный документ делает шаг в этом направлении – в нём впервые поднята проблема количественной оценки и изучения взаимодействия читателей Википедии с цитатами. Для большей конкретизации мы задаём следующие исследовательские вопросы:

Вопрос 1: Как часто пользователи переходят к цитатам при чтении Википедии? (Раздел 4.)

Вопрос 2: Какие особенности страницы предсказывают, будет ли читатель взаимодействовать с цитатой на странице? (Раздел 5.)

Вопрос 3: Какие особенности цитаты предсказывают, будет ли читатель взаимодействовать с этим материалом? (Раздел 6.)

Чтобы ответить на эти вопросы, мы собрали большой набор данных (объёмом 96 Мб), включающий все связанные с цитированием действия в англоязычной Википедии за два месяца (октябрь 2018 г. и апрель 2019 г.), в том числе справочные щелчки (клики), всплывающие справки, сноски (примечания) вниз и вверх, как это показано на рисунке.



Примеры шести типов взаимодействий со страницами и цитатами, которые мы фиксируем в английской Википедии с помощью инструмента *EventLogging* Викимедиа. Страница Википедии состоит из нескольких частей: заголовок статьи, основной текст статьи, зона (раздел) примечаний внизу основного текста, зона дополнительной информации в правой части страницы. Во всех зонах отдельные слова или несколько слов могут снабжаться активируемыми гиперссылками

Анализируя этот набор данных (данные доступны на https://github.com/epfl-dlab/wikipedia-citation-engagement), мы делаем следующие основные выводы.

Мы количественно оцениваем взаимодействие пользователей с цитатами и находим, что это относительно редкое событие (*RQ1*, раздел 4): в течение одного месяца 93% ссылок в цитатах никогда не ак-

тивируются, а доля сетевых страниц, с которых происходил переход по ссылке, составляет 0,29%.

Мы получаем представление о факторах, связанных с поиском дополнительной информации посредством использования цитирования, как на уровне страницы (раздел 5), так и на уровне ссылок (раздел 6). С помощью сопоставимых сравнений показываем, что статьи, которые имеют более высокое качество и, следовательно, более популярные и подробные, способствуют снижению склонности пользователей работать с цитатами (не стимулируют интерес читателей к цитированию). Применяя модель логистической регрессии, настроенную распознавать лингвистические особенности, мы определили, что чаще используемые ссылки, как правило, связаны с социальными или жизненными событиями.

Таким образом, мы приходим к выводу, что для читателей Википедия становится переходным мостиком (шлюзом) в тех случаях, когда сами статьи невысокого качества, либо недостаточно информативны. А если статьи в Википедии и информация в них достаточно высокого качества, то в подавляющем большинстве случаев Википедия является конечным пунктом назначения (не нужны разъяснения или дополнения).

В нашей работе впервые исследовано, взаимодействуют ли пользователи с цитатами в Википедии и каким образом; мы прокладываем путь к более широкому и глубокому пониманию роли Википедии в глобальной информационной экосистеме.

2. Смежные работы

Этот доклад связан с исследованиями по ряду смежных направлений (тем).

Характеристики читателей Википедии. Значительная часть предшествующих исследований была направлена на понимание работы пользователей с Википедией с позиций редакторского сообщества [**2**, **41**, **43**, **68**]. Исследования поведения читателей Википедии в основном учитывают интерес к содержанию (контенту) [**31**, **49**, **63**], популярность контента [**8**, **47**, **56**] или последовательность и время события [**37**]. Совсем недавно изучался вопрос, почему пользователи читают Википедию, комбинируя в опросы с несколькими вариантами ответов со статистическим анализом лог-файлов активности пользователей

[**53**]. Похожий по методике анализ касался изучения Википедии на 14 языках, кроме английского [**32**]. Мало известно, однако, о том, как пользователи взаимодействуют с цитатами Википедии из внешних источников; наше исследование – первое на эту тему.

Навигация по Википедии. Ссылочные материалы Википедии – часть обширной системы её связей с Сетью. Понимание того, как используются ссылки, может дать полезную дополнительную информацию о том, как улучшить функционирование этой связи [**29**, **66**]. В предыдущих исследованиях уже изучали, анализировали, моделировали и предсказывали базовые элементы навигации специалистов по Википедии [**13**, **18**, **21**, **30**, **52**, **62**], в основном опираясь на работу с игровыми материалами в *Wikispeedia* [**50**, **64**, **65**] и в *WikiGame* [**12**, **26**, **54**]. В нашем исследовании мы использовали новые, детально структурированные наборы данных по работе пользователей с научными ссылками из Википедии на внешние научные источники.

Наука в Википедии. Заметная часть ссылок в Википедии отсылает нас к научным публикациям [**40**]. Следовательно, она является базовым окном для научных результатов и позволяет широкой публике лучше понимать науку [**33**, **34**, **38**, **51**, **61**]. Вероятность цитирования конкретного материала в Википедии зависит от импакт-фактора публикации и доступности этого источника в режиме открытого доступа [**58**]. Сам факт цитирования в Википедии может рассматриваться как компонента импакт-фактора. Несмотря на то, что Википедия имеет косвенное влияние на ход научного прогресса [**59**], обратные ссылки на неё в научной литературе достаточно редки [**23**, **60**].

Совершенствование Википедии. Качество материалов Википедии зависит от работы редакторов и постоянного совершенствования текста статей [**9**, **46**]. Автоматические или полуавтоматические инструменты [**17**, **39**, **44**] помогают как совершенствовать пользовательский опыт [**29**, **69**], так и разнообразить содержание [**42**, **67**] и улучшить его качество [**1**, **20**, **28**]. Надёжность и достоверность материалов Википедии могут повышаться автоматически, например, посредством обнаружения материала для потенциальных ссылок [**15**] или обращения через Википедию с целью поиска материалов для цитирования [**48**]. Глубокое изучение нашей работы поможет улучшить Википедию путём добавления новых цитат, которые, возможно, привлекут пользователей.

Количественные оценки работы пользователей. Вовлечение пользователей, их реакция критически важны для совершенствования сервисов Википедии. Поэтому многочисленные исследователи сфокусировались на получении количественных оценок работы пользователей Сети с онлайновыми документами, например, в компьютерных рекламных процессах [**6**, **70**], социальных сетях [**5**, **10**, **22**] или при поиске информации [**24**, **55**].

3. Сбор данных о цитировании

Чтобы изучить работу читателей с цитатами, мы организовали сбор данных о том, как читатели ориентируются и как они взаимодействуют с цитатами в англоязычной Википедии.

Общая информация: цитаты в Википедии

Статьи в Википедии написаны редакторами в вики-коде, программе языка разметки, затем они переводятся в *HTML* программным обеспечением *MediaWiki*, которое обеспечивает функционирование сайта.

Есть разные способы добавить ссылки на источники в тексте, они кратко изложены ниже. Во всех случаях полные справочные описания представлены в виде примечаний (сносок) внизу страницы (в специальной зоне (разделе) под названием «Ссылки» (в русскоязычных текстах их называют примечаниями внизу страницы. – *Примеч. пер.*) с автоматически присвоенным номером сноски, добавляемым в качестве привязки (якоря) ссылки (например, «[**1**]») в тексте статьи везде, где цитируется эта ссылка (см. рис. выше). Большинство ссылок в этом разделе состоит из текста, включая название источника, имена авторов, год публикации и издателя источника.

Для 80% ссылок Википедии заголовок источника активируется посредством клика гиперссылки на источник (*титул кликабелен*). Кроме того, при чтении страницы, когда курсор проходит над номером примечания, возникает всплывающее окно, содержащее текст ссылки и кликабельную ссылку, если таковая имеется, например: https://www.mediawiki.org/wiki/Reference_Tooltips. Daniel Nasaw (July 24, 2012). “Meet the ‘bots’ that edit Wikipedia”. BBC News («Познакомьтесь с “ботами”, которые редактируют Википедию». Новости BBC). Когда читатели нажимают на номер сноски (ссылки, примечания), они переходят

на описание ссылки внизу страницы, откуда они могут вернуться назад к тем местам, нажав на маленькую иконку (например, ^).

Наиболее распространённый способ добавить ссылку на статью в соответствии с руководящими принципами Википедии – делать это с помощью встроенной ссылки, используя тег <*ref/*> непосредственно в контексте, где ссылка впервые цитируется. В теге редакторы могут указывать справочные данные (текст и ссылки), используя заранее предопределённый шаблон или простой вики-код. В дополнение к этому стандартному методу некоторые ссылки в инфобоксе добавляются автоматически по шаблонам, включённым в страницу, например таким, как геолокация. Стоит отметить, что ссылку можно приводить несколько раз, присвоив ей имя и добавив тег к каждому предложению, которое должно ссылаться на него. С учётом многочисленных способов использования тега <*ref/*> и для точного подсчёта в данной работе мы переводили текст статьи из вики-кода в *HTML* и извлекали информацию из кода (программы) *HTML*.

Регистрация случаев цитирования и загрузки страницы

Мы используем инструмент *EventLogging* Викимедиа (https://www.mediawiki.org/wiki/Extension:EventLogging/Guide) – расширение программного обеспечения *MediaWiki,* которое осуществляет ведение журнала лог-файлов на стороне клиента, фиксируя конкретные типы событий.

Мы отслеживаем пять основных типов действий, связанных с цитированием, а также шестое событие – «загрузка одной страницы». Для этого мы фиксируем действия мыши – любое взаимодействие читателя со ссылкой (см. рис. выше для визуального пояснения). Перечислим фиксируемые программой элементы:

1) само событие просмотра какой-либо определённой страницы Википедии, ему присвоено название «загрузка страницы» (*pageLoad)*;

2) щелчок левой кнопкой мыши (клик) по гиперссылке в основном тексте, который приводит пользователя в справочный раздел (к примечаниям) внизу страницы. Этому событию присвоено название *fnClick* (от *footnote*);

3) обратное по отношению к предыдущей операции действие: щелчок по номеру сноски в разделе «Примечания», который приведёт пользователя обратно в то место основного текста, откуда он отправился посмотреть справку. Этому событию присвоено название *upClick*;

4) щелчок левой кнопкой мыши (клик) по гиперссылке в разделе «Справки» («Примечания») в нижней части страницы. Этому событию присвоено название *refClick* (от *reference*);

5) клик по внешней ссылке, которая находится вне пределов основного текста статьи и раздела «Примечания», в специальной справочной зоне справа от статьи. Этому событию присвоено название *extClick* (от *external*);

6) приостановка курсора мыши (либо замедление движения) более чем на 1 с над гиперссылкой в тексте активирует всплывающую подсказку, которая видна не в разделе «Примечания», а непосредственно в том месте, где она активирована. Этому событию присвоено название *fnHover* (от *footnote* – примечание и *hover* – парить, всплывать. Подсказка *(tooltip, hint)* – элемент графического интерфейса, служит дополнительным средством обучения пользователя).

Платформа *EventLogging* управляет так называемым маркером сессии (сеансовым токеном), основанным на технологии *cookie*-идентификаторе, группирующем события, которые произошли в одной и той же закладке браузера. Таким образом, мы ссылаемся на последовательности событий, происходящих с одним и тем же маркером сессии. Мы собрали данные для мобильных устройств и настольных персональных компьютеров о трафике событий, связанных с цитированием за два периода по четыре смежных недели: с 26 сентября по 25 октября 2018 г. и с 24 марта по 21 апреля 2019 г.

В обоих случаях мы собрали все связанные с цитированием события (*extClick, refClick, fnHover, fnClick, upClick*) и (из-за ограничений, связанных с недостатком мощности вычислительной инфраструктуры) сделали выборку событий *pageLoad* на уровне сеанса в размере 33%.

Чтобы убедиться, что журналы отражают поведение читателя, а не поведение редактора, мы исключительно сохранили данные за эти четыре недели только от анонимных пользователей, отбрасывая все события, сгенерированные редакторами Википедии (вошедшие в систему пользователи или зарегистрированные пользователи). Также отбрасы-

вали данные от ботов (боты фильтруются детекторами – фильтрами, встроенными в программу *EventLogging*). Мы собирали данные ровно четыре недели, чтобы уменьшить потенциальные неточности из-за неравномерности частоты дней недели. На протяжении всей работы мы сосредоточились в основном на данных за второй период (апрель 2019 г.), а октябрьские данные 2018 г. использовали только для продольного (контрольного) исследования воздействия качества статьи на взаимодействие читателей с цитатами.

Определение индикаторов (метрик) взаимодействия

Два ключевых показателя в нашем анализе – это темп перехода по ссылкам (*click-through rate CTR*) и темп прохода по всплывающим подсказкам (примечаниям).

Пусть для каждой страницы *p* и каждой сессии *s C (p, s)* индикативная функция будет равна 1, если на странице была активирована (нажата) хотя бы одна ссылка во время сеанса *s* соответствующим пользователем (событие *refClick*), и будет равна 0 в противном случае – если этого не произошло.

Аналогичным образом, пусть *H (p, s)* указывает, что пользователь зашёл на хотя бы одну всплывающую сноску (событие *fnHover*). Далее, пусть *N (p)* – это количество сессий, в течение которых была загружена страница *p* (событие *pageLoad*).

Глобальный рейтинг кликов. Глобальный рейтинг кликов *CTR* отражает в целом работу читателя с помощью ссылочных кликов по Википедии. Определяется как доля просмотров страниц, на которых произошёл щелчок хотя бы по одной ссылке (при этом все просмотры одной и той же страницы за одну сессию считаются одним событием):

$$gCTR = \frac{\sum_p \sum_s C(p,s)}{\sum_p N(p)},$$

где *p* обозначает множество страниц, которые содержат хотя бы одно примечание с гиперссылкой.

Темп переходов для данной страницы определяется как вероятность наблюдения, по крайней мере, одного щелчка на ссылку во время сеанса, в котором была просмотрена страница *p*:

$$pCTR(p) = \frac{\sum_s C(p,s)}{N(p)}.$$

Наконец, мы обозначаем средний для страницы *CTR* для множества страниц *P* как

$$pCTR(p) = \frac{1}{|P|}\sum_{p \in P} pCTR(p).$$

Обратите внимание, что средний для страницы *pCTR (P)* соответствует макросреднему значению, в котором каждая страница имеет один и тот же вес, тогда как глобальный *gCTR* соответствует микросредним значениям, где страницы взвешены пропорционально количеству сессий, в которых они были просмотрены.

Исследование сносок. По аналогии с приведёнными выше определениями, но заменив индикатор щелчка *C (p, s)* на индикатор всплывающих ссылок *H (p, s)* мы получаем глобальную и специфическую для страницы величину темпа всплывания ссылок.

$$gHR = \frac{\sum_p \sum_s H(p,s)}{\sum_p N(p)}, \qquad pHR(p) = \frac{\sum_s H(p,s)}{N(p)}.$$

Фиксация (захват) контекста события. Каждое событие характеризуется набором особенностей, которые фиксируют информацию о трёх аспектах события:

сеанс (сессия), в котором произошло событие;

страница;

ссылка.

Сессия – мы собираем уникальный токен сессии (закладку, *cookie,* см. раздел 3.2), который идентифицирует закладка браузера, где произошло событие.

Страницы – на уровне статьи мы храним заголовок, идентификатор страницы, длину текста вики-кода в символах, количестве ссылок и

популярность (число событий *pageLoad* в период сбора данных). Мы также используем общий (черновой) классификатор (*drafttopic classifier*) *ORES* [**3**] для сопоставления каждой статьи в Википедии с вектором тематик, элементы которых отражают вероятность того, чтобы страница соответствовала одной из 44 тем самого высокого уровня *WikiProjects* таксономии (см.: 7https://en.wikipedia.org/wiki/Wikipedia: WikiProject_ Council/Directory).

Далее мы используем модель качества товара *ORES* [**20**] для маркировки изделий с уровнем качества, который может принимать следующие значения (от низкого до высокого качества): «Отбросы, Затычка», «Начальный уровень», «С-класс», «В-класс», «Хорошая статья», «Рекомендованная статья».

Ссылки – для каждой ссылки, нажатой или всплывшей, мы записываем ее *URL;* текст в ссылке; текст предложения, в котором ссылка указана; её относительную позицию (в какой части страницы она находится) на той странице, где ссылка цитируется. Так как мы связываем ссылки с контекстом, ссылки из одного источника, взятые на разных страницах, рассматриваются как отдельные.

Википедия динамична по своей природе: статьи постоянно обновляются, и их изменения отслеживаются путём пересмотра. К аккаунту для развития статей за четыре недели сбора данных мы присоединяем отдельные показатели на уровне статьи. Чтобы вычислить характеристики статьи, такие как длина статьи или количество ссылок, мы рассчитываем их среднее значение по всем просмотрам, начиная с периода регистрации. Чтобы определить уровень вовлечённости читателей по отношению к данной статье (например, загрузки страницы, клики по ссылкам), мы суммируем все события, которые записываются при каждом просмотре статьи.

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

Перевод А. И. Земскова, ГПНТБ России

(Продолжение в следующих номерах журнала.)

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

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 1)

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.

ACM Reference Format:

Tiziano Piccardi, Miriam Redi, Giovanni Colavizza, and Robert West. 2020. Quantifying Engagement with Citations on Wikipedia. In Proceedings of The Web Conference 2020 (WWW'20), April 20–24, 2020, Taipei, Taiwan. ACM, New York, NY, USA. 12 pages. https://doi.org/10.1145/3366423.3380300.

1. Introduction

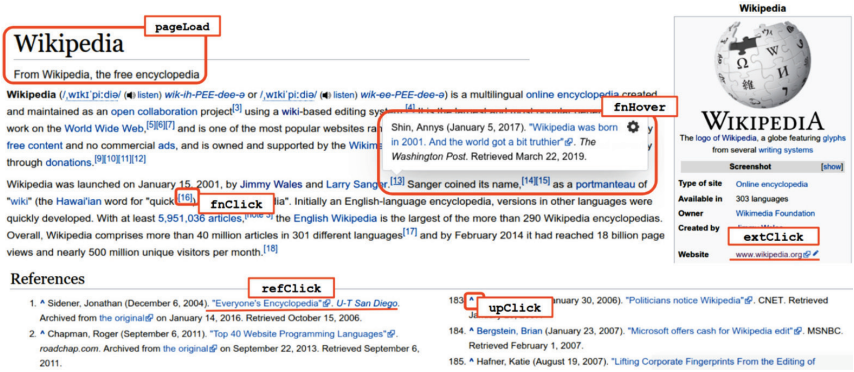
Wikipedia is the largest encyclopedia ever built, established through the collaborative effort of a large editor base, self-governed through agreed policies and guidelines [**7, 16**]. Thanks to the tenacious work of the editor community, Wikipedia's content is generally up to date and of high quality [**25, 45**], and is relied upon as a source of neutral, unbiased information [**35**].

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.

WWW'20, April 20–24, 2020, Taipei, Taiwan.

2020 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.

ACM ISBN 978-1-4503-7023-3/20/04. https://doi.org/10.1145/3366423. 3380300.



Examples of the 6 types of interactions with pages and citations that we record on English Wikipedia using Wikimedia's EventLogging tool

Wikipedia's inline references, or citations[1], are a key mechanism for monitoring and maintaining its high quality. Wikipedia's core content policies require that "people using the encyclopedia can check that the information comes from a reliable source"[2] and citations are the main way to connect a statement to its sources. A clearly distinctive feature of Wikipedia is the fact that many citations are actionable: they are often equipped with hyperlinks to the cited material available on the Web.

As a result, Wikipedia's role on the Web has been defined as the "bridge to the next layer of academic resources" [**19**], and the "gateway through which millions of people now seek access to knowledge" [**11**]. Nevertheless, a question remains open: to which extent do Wikipedia readers actually cross the bridge and access the broader knowledge referenced in the encyclopedia?

Given the collaborative and open nature of Wikipedia, being able to quantify readers' engagement with the content and its supporting sources is of crucial importance for the constant betterment of the encyclopedia

---

1 We use the terms "reference" and "citation" largely interchangeably.

2 https://en.wikipedia.org/wiki/Wikipedia:Verifiability, https://en.wikipedia.org/wiki/Wikipedia:Reliable_sources.

and its role in fostering a self-critical society. By understanding readers' interactions with citations, we can better assess the role of Wikipedia editors and policies in maintaining a high quality of information, measure public demand for secondary sources, and provide insights and potential recommendations to increase the public's interest in references.

This paper takes a step in this direction, by addressing, for the first time, the problem of quantifying and studying Wikipedia readers' engagement with citations. More specifically, we ask the following research questions:

RQ1. To what extent do users engage with citations when reading Wikipedia? (Sec. 4.)

RQ2. What features of a page predict whether a reader will interact with a citation on the page? (Sec. 5.)

RQ3. What features of a citation predict whether a reader will interact with it? (Sec. 6.)

In order to answer these questions, we collect a large dataset comprising all citation-related events (96M) on the English Wikipedia for two months (October 2018, April 2019), including reference clicks, reference hovers, and downwards and upwards footnote click, as visualized in figure. By analyzing this dataset[3], we make the following main contributions.

We quantify users' engagement with citations and find that it is a relatively rare event (RQ1, Sec. 4): 93% of the links in citations are never clicked over a one-month period, and the fraction of page views that involve a click on a citation link is 0,29%.

We gain insights into factors associated with seeking additional information via citation interactions, both at the page level (RQ2, Sec. 5) and at the link level (RQ3, Sec. 6). Through matched observational studies, we show that articles that are of higher quality, and thus also longer and more popular, are associated with a lower propensity of users to interact with citations. Using a logistic regression model trained on linguistic features, we show that more frequently clicked citation links tend to relate to social or life events.

---

[3] Notebooks with code at https://github.com/epfl-dlab/wikipedia-citation-engagement.

We thus conclude that readers are more likely to use Wikipedia as a gateway on topics where Wikipedia is still wanting and where articles are of low quality and not sufficiently informative; and that Wikipedia tends to be the final destination in the large majority of cases where the information it contains is of sufficiently high quality.

Our work provides the first study aimed at understanding if and how users engage with citations on Wikipedia, thus paving the way for a broader and deeper understanding of Wikipedia's role in the global information ecosystem.

2. Related work

This paper is related to research on a number of different themes.

Characterizing Wikipedia readers. A substantial amount of prior work has focused on understanding user engagement with Wikipedia from the point of view of the editor community [**2, 41, 43, 68**]. Studies on the behavior of Wikipedia readers have mostly considered interest in contents [**31, 49, 63**], content popularity [**8, 47, 56**], or event timing [**37**]. More recently, a study explored the question why users read Wikipedia, by combining multiple-choice surveys with log-based analyses of user activity [**53**]. A similar design was used to study 14 languages other than English [**32**]. Little is known, however, on how users engage with Wikipedia's citations of external sources; ours is the first study on this subject.

Navigation in Wikipedia. Wikipedia citations are part of the hy-perlink network connecting Wikipedia and the Web. Understanding citation usage can yield useful insights for improving this network [**29, 66**]. The analysis, modeling, and prediction of human navigation inside Wikipedia has been considered in previous studies [**13, 18, 21, 30, 52, 62**], largely relying on traces from the navigation games Wikispeedia [**50, 64, 65**] and WikiGame [**12, 26, 54**]. For our study, we collect instead a new, fine-grained dataset of user interactions with Wikipedia references to external content.

Science in Wikipedia. A sizeable portion of citations on Wikipedia refer to scientific literature [**40**]. Consequently, Wikipedia is a fundamental gateway to scientific results and enables the public understanding of science [**33, 34, 38, 51, 61**]. The chance of a scientific reference being cited on Wikipedia varies with the impact factor of the publication venue and its open-access availability [**58**]. Being cited on Wikipedia can thus be considered an indicator of impact [**27**]. Despite the indirect influence that Wikipedia has on scientific progress [**59**], Wikipedia is in turn rarely acknowledged in the scientific literature [**23, 60**].

Improving Wikipedia. Wikipedia content quality relies on the work of editors and their gradual improvement of articles [**9, 46**]. Automated or semiautomated tools [**17, 39, 44**] can help improve user experience [**29, 69**], content variety [**42, 67**], and quality [**1, 20, 28**]. The reliability of Wikipedia can also be improved automatically, e.g., by finding potential citations [**15**] and Wikipedia statements in need of evidence [**48**]. The insights from our work can help improve Wikipedia via new citations with which users would be more likely to interact.

Quantifying Web user engagement. User engagement is crucial for the success of Web services, and numerous researchers have focused on quantifying how Web users engage with online content, e.g., in computational advertising [**6, 70**], social media [**5, 10, 22**], or information retrieval [**24, 55**]. Also, while the body of work focusing on understanding readers' and editors' engagement with content within Wikipedia has been growing in the recent years [**36**], we study here for the first time how Wikipedia readers engage with the broader outside knowledge linked from the online encyclopedia.

3. Citation data collection

To study readers' engagement with citations, we collected data capturing where readers navigate and how they interact with citations in English Wikipedia.

3.1. Background: Citations in Wikipedia

Articles in Wikipedia are written by editors in wikicode, a markup language that is then translated to HTML by MediaWiki, the software that powers the website. There are different ways to add citations to sources in the text, summarized below. In all cases, the full reference descriptions are rendered as footnotes at the bottom of the page (in a dedicated section called References) with an automatically assigned footnote number that is added as a link anchor (e.g., [**1**]) in the text of the article wherever the reference is cited (figure). Most references in the References section consist of text including the title of the source, the authors' names, the year of publication, and the source's publisher. For 80% of Wikipedia references, the source title is actionable via a clickable link to the source. Also, when reading a page, hovering over a reference's footnote number with the mouse cursor will display a reference tooltip[4], a pop-up containing the reference text and a clickable link (when present), e.g., Daniel Nasaw (July 24, 2012). "Meet the 'bots' that edit Wikipedia". BBC News.

When readers click on the reference's footnote number, they are sent to the reference description at the page bottom, from where they can jump back to the locations where the reference is cited by clicking on a small icon (e.g., ^).

The most common method to add a reference to an article, also recommended by the Wikipedia guidelines, is via an inline citation using a <ref/> tag directly in the context where the reference is first cited. In the tag, the editors can specify the reference details (text and links) by using a predefined template or plain wikicode. In addition to this standard method, some references are added automatically by templates included in the page, such us the geolo-cations present in the infobox. It is worth noting that a reference can be cited multiple times by assigning it a name and appending the tag to every sentence that should link to it. Given the numerous ways to use the <ref/> tag, and in order to have an accurate view of the article, we parsed pages from wikicode to HTML and extracted the information from the HTML code.

---

[4] https://www.mediawiki.org/wiki/Reference_Tooltips.

3.2. Logging citation and page load events

We make use of Wikimedia's EventLogging tool[5], an extension of the MediaWiki software that performs client-side logging of specific types of events. We detect 5 main types of citation-related events and 1 page load event. In terms of citations, we capture the mouse events that involve any kind of reader interaction with the references (see figure for a visual explanation):

refClick: a click on a hyperlink in an article's reference section.

extClick: a click on an external link outside the reference section.

fnHover: a hover over a footnote number in the text, logged when the reference tooltip is visible for more than 1 second.

fnClick: a click on a footnote number, which takes the user to the reference section at the bottom of the page.

upClick: the inverse of fnClick: a click on a reference's up arrow icon that takes the reader back to the part of text where the reference is cited.

pageLoad: in addition to the above citation-related events, this event is triggered whenever a Wikipedia article is loaded.

The EventLogging platform manages a so-called session token, a cookie-based identifier that allows us to group events that happened within the same browser tab. We henceforth refer to event sequences that occur with the same session token as sessions.

We collected 4 contiguous weeks[6] of Wikipedia mobile and desk-top traffic data of citation-related events. We repeated the 4-week data collection over two periods: from September 26 to October 25, 2018, and from March 24 to April 21, 2019. In both cases, we collected all citation-related events (extClick, refClick, fnHover, fnClick, upClick) and (due to computational infrastructure constraints) sampled pageLoad events at the session level at a rate of 33%.

---

[5] https://www.mediawiki.org/wiki/Extension:EventLogging/Guide.

[6] We collected exactly 4 weeks to reduce potential seasonal effects due to uneven day-of-the-week frequencies.

To ensure that the logs reflect reader, rather than editor, behavior, we exclusively retained data from users who in the 4 weeks of data collection acted only as anonymous readers, discarding all events generated by Wikipedia editors (logged in users or users with anonymous edits) and by bots (which can be filtered out using a detector provided by the EventLogging tool).

Throughout the paper, we will mostly focus on the data from the second data collection period (April 2019) and only use the October 2018 data for a longitudinal study measuring the impact of article quality on readers' engagement with citations.

We collected exactly 4 weeks to reduce potential seasonal effects due to uneven day-of-the-week frequencies.

3.3. Definition of engagement metrics

Two key metrics in our analysis will be the citation click-through rate (CTR) and the footnote hover rate.

For each page p and each session *s*, let *C (p, s)* be the indicator function that is 1 if at least one reference was clicked on page *p* during session *s* by the respective user (refClick event), and 0 otherwise. Analogously, let *H (p, s)* indicate if the user hovered over at least one footnote (fnHover event). Furthermore, let *N (p)* be the number of sessions during which *p* was loaded (page load event).

Global click-through rate. The global CTR measures overall reader engagement via reference clicks across Wikipedia. It is defined as the fraction of page views on which at least one reference click occurred (treating all views of the same page in the same session as one single event):

$$gCTR = \frac{\sum_p \sum_s C(p,s)}{\sum_p N(p)},$$

where *p* ranges over the set of pages that contain at least one reference with a hyperlink.

Page-specific click-through rate. The page-specific *CTR* for page *p* is defined as the probability of observing at least one click on a reference in *p* during a session in which p was viewed:

$$pCTR(p) = \frac{\sum_s C(p,s)}{N(p)}.$$

Finally, we denote the average page-specific *CTR* over a set *P* of pages by

$$pCTR(p) = \frac{1}{|P|}\sum_{p \in P} pCTR(p).$$

Note that *pCTR(P)* corresponds to a macro average where every page gets the same weight, whereas *gCTR* corresponds to a micro average where pages are weighted in proportion to the number of sessions in which they were viewed.

Footnote hover rates. In analogy to the above definitions, but when replacing the click indicator *C (p, s)* with the hover indicator *H (p, s)*, we obtain the global and page-specific footnote hover rates:

$$gHR = \frac{\sum_p \sum_s H(p,s)}{\sum_p N(p)}, \qquad pHR(p) = \frac{\sum_s H(p,s)}{N(p)}.$$

3.4. Capturing event context

Each event is characterized by a set of features that capture information about three aspects of the event: the session in which the event happened, the page, and the reference.

Session: We collect the unique session token (cf. Sec. 3.2) that identifies the browser tab in which the event occurred.

Pages. At the article level, we store title, page id, text length of wikicode in characters, number of references, and popularity (number of pageLoad events during the data collection period). We also use the ORES drafttopic classifier [**3**] to label each Wikipedia article with a vector of topics, whose elements reflect the probability of the page to belong to one the 44 topics from the highest level of the WikiProjects taxonomy[7]. We further use the ORES articlequality model [**20**] to label articles with a quality level, which can take the following values (from low to high quality): “Stub”, “Start”, “C-class”, “B-class”, “Good Article”, “Featured Article”.

References. For each reference clicked or hovered, we record its URL, the text in the reference, the text of the sentence in which the reference is cited, and the relative position (character offset from the start in plain text, divided by page length) in the page where the reference is cited. Since we associate references to their contexts, references to the same source appearing on different pages are treated as distinct.

Wikipedia is dynamic by nature: articles are continuously updated, and their changes are tracked through revisions. To account for the evolution of articles over the 4 weeks of data collection, we aggregate individual revision-level metrics at the article level. To compute article-specific characteristics such as article length or number of references, we calculate their average over all revisions from the logging period. To quantify the amount of reader engage-ment with a given article (e.g., page loads, reference clicks), we sum all events recorded at each revision of the article.

[7] https://en.wikipedia:WikiProject_Council/Directory.

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