DOI: 10.33186/1027-3689-2020-10-63-76

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 2)

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 года, 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.

Общая статистика англоязычной Википедии

К моменту завершения работы по сбору данных англоязычная Википедия содержала 5,8 млн статей, 5,4 млн (95%) из которых при подготовке наших данных были загружены по крайней мере один раз, в общей сложности состоялось 7,4 млн просмотров.

Из просмотренных статей 3,9 млн (73%) содержат по крайней мере одну ссылку, всего система ссылается на 24 млн различных URL-адресов.

За 4 недели работы по сбору данных мы собрали (при объёме выборки 33%) 1,5 млрд событий *pageLoad* (из них 62% выгружено с помощью мобильных устройств и остальные – с рабочего стола ПК).

На рис. 2а показано нарастающим итогом (дополнительное кумулятивное) распределение популярности для страниц Википедии, которые были просмотрены хотя бы один раз за период сбора данных.

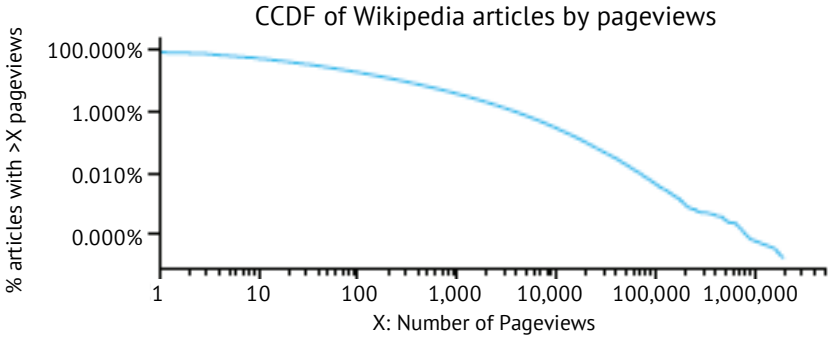

Рис. 2a. Распределение статей Википедии по популярности (количество просмотров страниц; комплементарная интегральная функция распределения – *Complementary Cumulative Distribution Function, CCDF;* горизонтальная ось – количество просмотров статей в логарифмическом масштабе; вертикальная ось – доля статей с соответствующим количеством просмотров)

Распределение сильно искажено, примерно 83% статей загружалось менее 100 раз в 33% случайной выборки или менее 300 раз при экстраполяции результатов на все данные.

Мы наблюдаем аналогичное неравномерное распределение длины страницы (рис. 2b), причём большинство статей очень короткие.

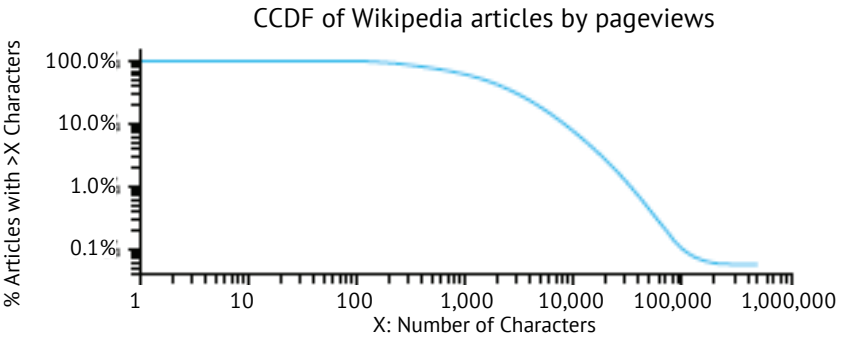

Рис. 2b. Распределение статей Википедии по длине страницы (количество символов в программе *wikicode*; горизонтальная ось – количество символов в статье в логарифмическом масштабе; вертикальная ось – доля статей с соответствующим количеством символов (комплементарная интегральная функция распределения)

На рис. 2c показано, что распределение уровней качества статей также сильно искажено в сторону низкого уровня качества: большинство статей определяется как «Огрызок» или «Начальный уровень», и менее 300 тыс. статей помечены как «Хорошие» или «Рекомендованные».

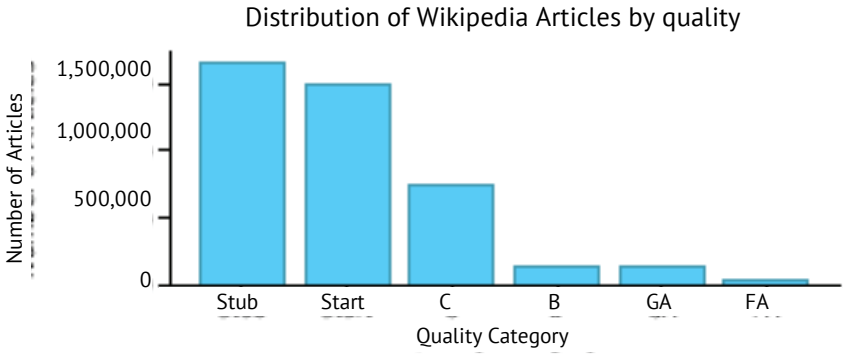

Рис. 2c. Распределение статей Википедии по категориям качества (горизонтальная ось – качество увеличивается слева направо; *Stub* – «Отбросы, затычка», *Start* – «Начальный уровень», *C* – «С-класс», *B* – «В-класс», *GA* – «Хорошая статья», *FA* – «Рекомендованная статья»)

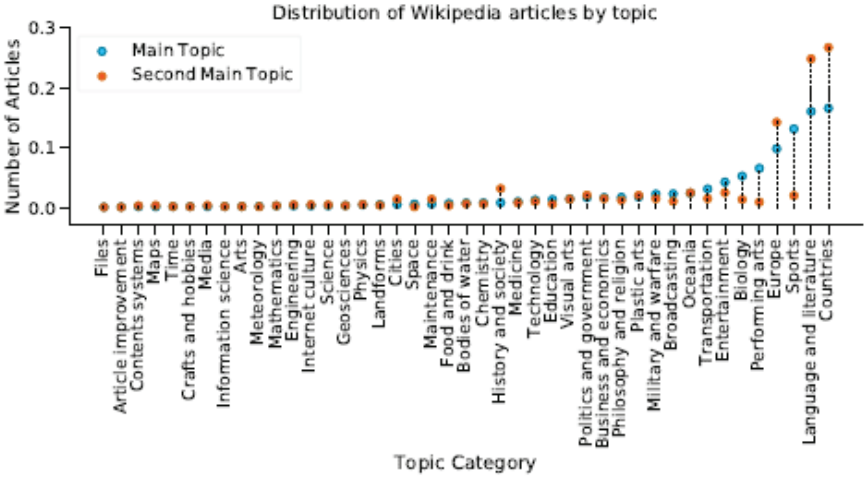

Рис. 3. Распределение тематики статей в Википедии: наиболее популярные тематики – голубые точки, вторые по популярности – оранжевые точки

Мы обнаружили, что большинство статей посвящено географии или тематике «Язык и литература» (последняя включает биографии), затем следуют темы, связанные со спортом и наукой (рис. 3).

4. Распространённость использования цитат

После вступительных обсуждений мы готовы обратиться к нашему первому научному вопросу «Как часто пользователи переходят к цитатам при чтении Википедии?» (раздел 4).

5. Распределение типов взаимодействия

Мы начали с анализа относительной частоты различных видов цитирований. За месяц сбора данных мы зафиксировали 96 млн случаев цитирования. На рис. 4 показано, как эти случаи распределяются по пяти типам событий с разбивкой по устройствам (мобильное устройство или рабочий стол ПК). Большинство взаимодействий со ссылками происходит на настольном компьютере, а не на мобильных устройствах, несмотря на то, что большинство загрузок страниц (62%) производится с мобильных устройств.

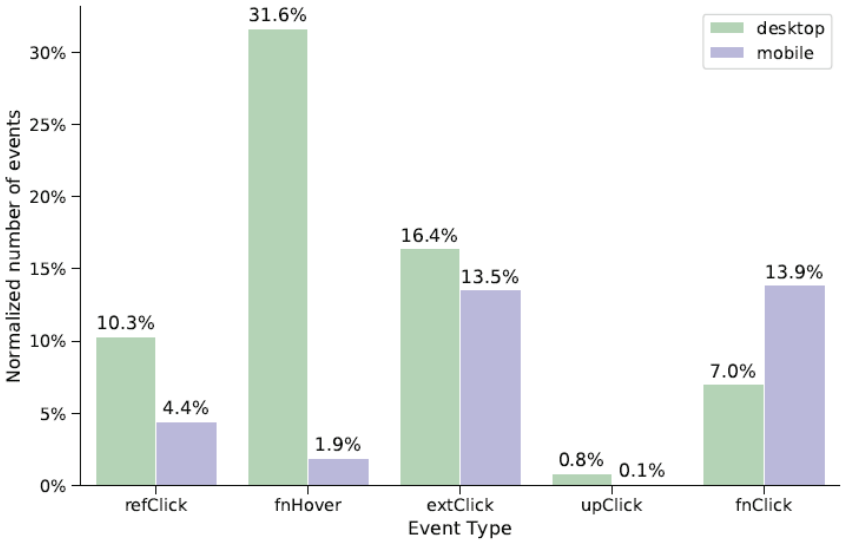

Рис. 4. Относительная частотность различных типов цитирования при работе с ПК (зелёные столбики) и при работе с мобильным устройством (голубые столбики) в апреле 2019 г. (по горизонтальной оси – тип события, по вертикальной оси – нормированное количество событий)

Взаимодействия также по-разному распределяются для мобильных устройств и для рабочего стола ПК. Наиболее распространённое событие при использовании рабочего стола – работа со всплывающей подсказкой (*fnHover*) для отображения справочного текста. Для активации всплывающей сноски требуется мышь, которая недоступна на большинстве мобильных устройств, что, в свою очередь, объясняет низкую частоту использования метода *fnHover* на мобильных устройствах.

Чтобы раскрыть текст ссылки за сноской, пользователям мобильных устройств нужно нажать на сноску, которая предположительно объясняет, почему *fnClick* является наиболее распространённым событием на мобильном телефоне.

Нажатие на вызов внешних ссылок за пределами раздела «Ссылки» в нижней части страницы (*extClick*) является вторым наиболее распространённым событием как на настольном, так и на мобильном уст-

ройстве, а затем по частотности следует нажатие на ссылки в нижней части страницы (тип ссылок *refClick*).

Наконец, действие *upClick,* которое позволяет пользователю перейти снизу из зоны (раздела) «Примечания, ссылки» в то место, где цитата инициировалась в основном тексте, почти никогда не применяется.

Темп перехода кликов

Мы сосредоточимся на двух наиболее распространённых взаимодействиях с цитатами: всплывающие ссылки (*fnHover*) и переход из основного текста в раздел «Примечания» нажатием по ссылкам цитирования (*refClick*). (Мы не останавливаемся на событиях *extClick,* так как они не касаются внутренних цитат, а относятся к внешним ссылкам.)

Во-первых, отметим, что из 24 млн различных предлагаемых к цитированию (активации) во всех статьях английской Википедии URL-гиперссылок 93% ни разу не были активированы во время месяца сбора данных.

Далее отметим, что общий темп кликов (*CTR*) по всем страницам с хотя бы одной ссылкой (глобальный *gCTR,* формула 1) составляет 0,29%, т.е. нажатия на ссылки происходят реже, чем 1 раз на 300 страниц. В анализе по типу устройства мы снова наблюдаем существенные различия между настольным компьютером и мобильным устройством: на настольном компьютере глобальный рейтинг кликов составляет 0,56%, что более чем в 4 раза выше, чем на мобильном телефоне, где он составляет всего 0,13%.

Средний *CTR* для конкретной страницы (*pCTR,* формула 3) несколько выше, он составляет 1,1% для настольных компьютеров и 0,52% для мобильных устройств. Это связано с тем, что там много редко просматриваемых страниц (см. рис. 2а) с высоким *CTR*. После исключения страниц с количеством просмотров менее 100 глобальный *CTR* составляет 0,67% для настольных компьютеров и 0,21% для мобильных устройств. Темп всплывающих сносок немного выше, глобальная величин темпа всплывания ссылок (*gHR,* формула 4) составляет 1,4%.

Средний для конкретной страницы темп всплывающей сноски (*pHR,* уравнение 4) составляет 0,68% при учёте всех страниц, по крайней мере с одной кликабельной ссылкой и 1,1% при исключении страниц, получивших (имеющих) менее 100 просмотров. Функция всплывания подсказок (ссылок) недоступна на большинстве мобильных уст-

ройств, поэтому цифры всплывающих ссылок относятся только к настольным устройствам.

В итоге мы отмечаем, что взаимодействие читателей с цитатами в целом низкое.

Влияние положения ссылки на странице

Ранее было показано, что пользователи Википедии чаще активируют внутренние гиперссылки на используемую литературу, которые расположены в верхней части страницы [**42**]. Чтобы проверить, верно ли это также и для ссылок, с которыми мы работаем, берём одну случайную загрузку страницы с цитированием за сеанс и случайным образом определённым одним кликом, а также одну ссылку без клика для этой же загрузки страницы. Затем определяем относительную позицию каждой ссылки на странице как смещение от верхней части страницы, делённое на длину страницы (в символах). Рис. 5, на котором показано относительное положение мест, где произошёл клик, и страниц без кликов, свидетельствует, что пользователи более часто нажимают на ссылки в верхней части страницы и не столь часто в нижней.

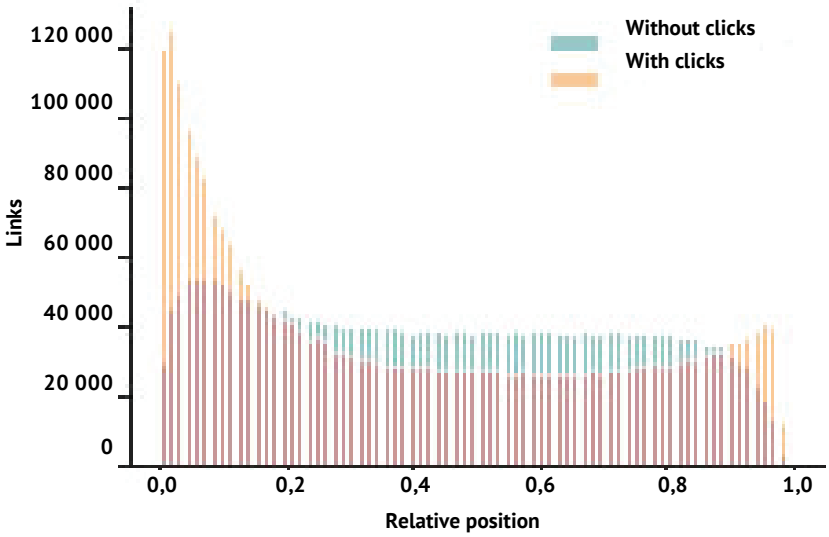

Рис 5. Относительное местоположение на странице Википедии задействованных ссылок (коричневые столбики) и незадействованных (голубые столбики)

Самые популярные домены

Посмотрим, на какие домены чаще всего переходят пользователи. На первых порах казалось, что чаще других посещается домен *archive.org* (интернет-архив) – 882 тыс. событий *refClick*. Такие URL-адреса обычно представляют собой снимки (снэп-шоты) старых веб-страниц, заархивированных в системе интернет-архив программой *Wayback Machine*. Поэтому для уточнения мы извлекаем исходные домены из архивной оболочки.

На рис. 6 представлены 15 наиболее востребованных доменов по количеству *refClick*. Самым популярным оказался *google.com*. При более детальном обследовании мы выявили, что значительная часть переходов ведёт на *books.google.com*, который обеспечивает частичный доступ к печатным источникам. Второй домен с наибольшим количеством ссылок – *doi.org* – для идентификации всех научных статей, отчётов и наборов данных, записанных с цифровым идентификатором объекта (*DOI*); затем следуют газеты (в основном либеральные: *The New York Times, The Guardian* и др.) и радиовещательные каналы (*BBC*).

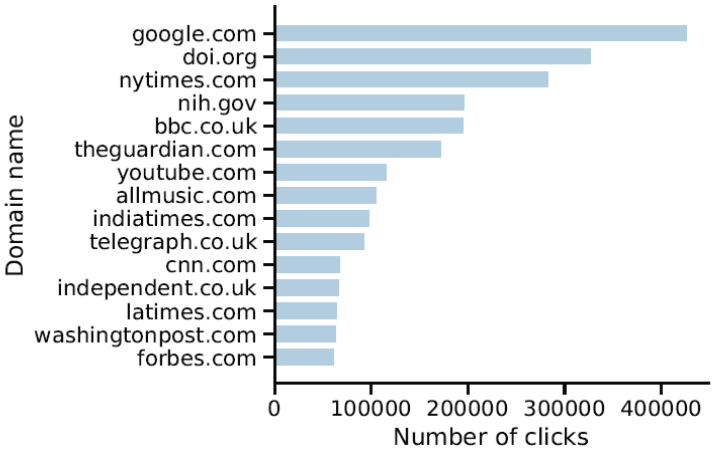

Рис. 6. Наиболее востребованные имена доменов в англоязычной Википедии (количество кликов за апрель 2019 г.)

(Сверху вниз: Гугл, *doi,* газета «Нью-Йорк Таймс» (*New York Times*), Национальный институт здоровья (*National Institute of health NIH*), программа *BBC*, газета *The Guardian*, система *YouTube*, *Allmusic* – онлайновая музыкальная база данных, газета «Таймс оф Индиа» (одна из самых читаемых и авторитетных газет Индии, по тиражу обходит все англоязычные крупноформатные газеты в мире), газета *Telegraph*, канал *CNN*, газета *Independent*, газета *Los Angeles Times*, газета *Washington Post, с*истема *Forbes*.)

Марковский анализ[*] цитирующих взаимодействий

Вышеприведённый анализ касался отдельных событий, а теперь попытаемся изучать сессии – это последовательность событий, которые произошли в той же закладке браузера (как указано в маркере сеанса). Каждая сессия начинается с события *pageLoad,* и мы добавляем специальный знак «*END* событие» после последнего фактического события в каждой сессии.

Подсчитывая переходы событий в сессиях, мы строим цепь Маркова первого порядка, задающую вероятность наблюдения *P (j | i)* события *j* сразу же после события *i,* где *i* и *j* могут принимать значения из того набора событий, который перечислен в разделе 3 (*pageLoad, refClick, extClick, fnClick, upClick, fnHover*) плюс специально введённое новое событие *END*.

Матрицы вероятности перехода для настольных компьютеров и для мобильных устройств приведены на рис. 7. Мы видим, что подавляющее большинство сеансов чтения состоит только из просмотров страниц – как на настольных, так и на мобильных устройствах; после загрузки страницы читатели склонны заканчивать сеанс (с вероятностью около 50%) или загрузить другую страницу в той же закладке (47%). Все свя-

---

[*] Марковский анализ – это метод, используемый для предсказания величины какой-либо переменной, если эта величина определяется только её нынешним (текущим) состоянием, а не какой-либо предшествовавшей активностью. По сути, этот метод предсказывает величину случайной переменной только на основе окружающих обстоятельств. – *Примеч. пер.*

занные с цитированием события имеют очень низкую вероятность (не более 1,2%) возникновения сразу после загрузки страницы.

При использовании рабочего стола ссылочные клики становятся намного более вероятными после кликов на сноски (34%), а клики сносок, в свою очередь, становятся значительно более вероятными при прохождении зоны всплывающих примечаний (6,5%), предрекая общий трёхшаговый сценарий (*fnHover, fnClick, refClick*), при котором читатель все глубже работает с цитатой. Обратите внимание, однако, что это неверно для мобильных устройств, где даже после того, как читатель нажал на сноску, вероятность, что он нажмёт на цитату, остаётся низкой (0,5%).

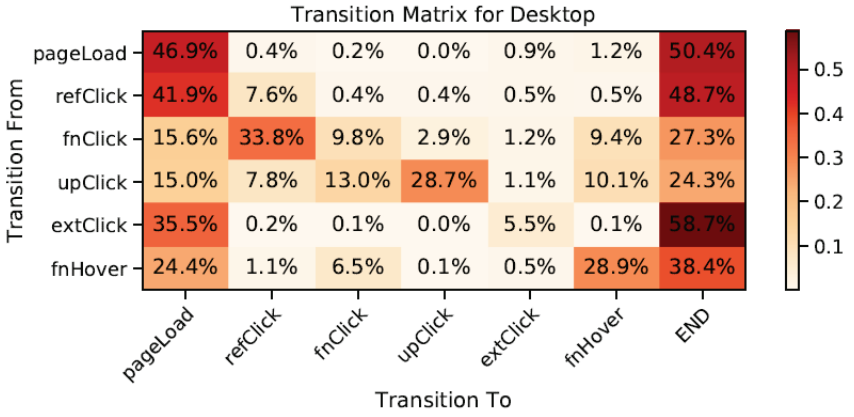

Рис. 7а. Поведение читателя, использующего настольный ПК. Матрица вероятностей переходов по цепи Маркова первого порядка от какого-либо события (*transition from*) к другому событию (*transition to*)

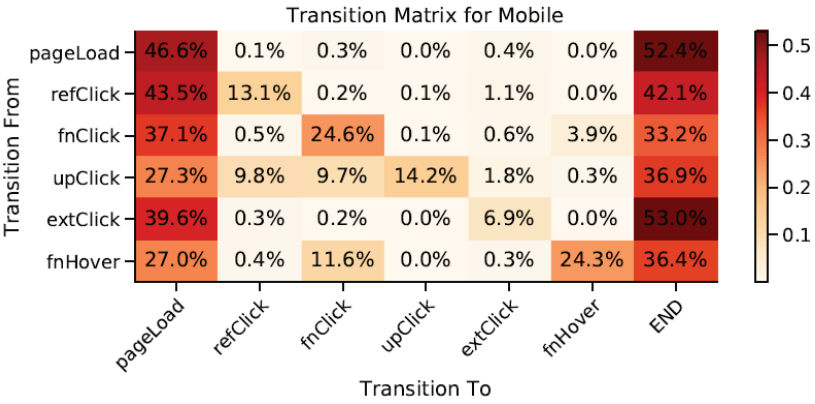

Рис. 7b. Поведение читателя, использующего мобильное устройство. Матрица вероятностей переходов по цепи Маркова первого порядка от какого-либо события (*transition from*) к другому событию (*transition to*)

Наконец, ссылочные клики (*refClick*) также распространены сразу после других ссылочных кликов (8% на рабочем столе, 13% на мобильном телефоне). Заметим, что для внешних ссылок вне раздела *References (extClick*) мы увидим другую картину: такие внешние клики редко следуют за взаимодействием с цитатами (*fnHover, fnClick, refClick*) и чаще всего (59% на настольных компьютерах, 53% на мобильных устройствах) они завершают сеанс, указывая на то, что Википедия в этих случаях обычно используется в качестве шлюза для выхода на внешние сайты.

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

(Продолжение в следующих номерах журнала.)

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