УДК 004.89 + 004.85 + 004.55 + 519.878
ББК 32.973.202

АНАЛИЗ СТРУКТУРЫ СЕТИ ИНТЕРНЕТ С ПОМОЩЬЮ ОБОБЩЁННЫХ МАРШРУТОВ

Тихонов А. В.[1]
(*Яндекс, Москва*)

Поисковые системы – важные элементы сети Интернет, помогающие пользователю находить нужную информацию. Развитие поисковых систем требует понимания задач пользователей сети Интернет и принципов их навигационного поведения. Для анализа навигационных паттернов предлагается понятие обобщённых маршрутов, с помощью которого проводится исследование страниц сети с точки зрения навигационного поведения пользователей по отношению к ним; выявляется несколько ярко выраженных классов страниц и демонстрируется, что, помимо схожих навигационных сценариев, страницы этих классов имеют также характерную специфику функциональности и содержимого. Исследование поведенческих паттернов, применяемых пользователями для решения задач, позволяет обнаружить сценарии навигации, которые сейчас плохо поддержаны, но предоставляют интерес для дальнейшего развития сервисов сети.

Ключевые слова: анализ сети Интернет, пользовательское поведение, навигация в сети Интернет, кластеризация страниц сети Интернет.

1. Введение

В настоящее время поисковые системы играют значительную роль в организации и поддержке пользовательского доступа к информации в сети Интернет по причине чрезвычайного

[1] *Алексей Тихонов, руководитель службы аналитики поиска, Яндекс, Москва* (*altsoph@yandex-team.ru*).

размера и сложности последней. Для предоставления ответов на пользовательские запросы поисковая система должна регулярно находить и анализировать структурированные и неструктурированные источники информации и затем генерировать ответы, представляемые на странице результатов поиска. Существует, однако, разница между представлением релевантного запросу результата на странице результатов поиска и предоставлением пользователю прямой ссылки на действительно нужную ему страницу, соответствующую решению его поисковой задачи. Одним из этапов решения задачи управления поисковыми системами и их совершенствования является подзадача выявления и анализа устойчивых пользовательских сценариев использования сети Интернет, поиска и потребления онлайн-контента.

Известно, что в настоящее время значительная часть трафика генерируется без какого-либо участия поисковых систем. Так, в [13] показано, что только 6,8% пользовательских сессий начинаются с поисковой системы, а в [11] указывается, что доля непосредственных визитов с поиска на типовой сайт в среднем составляет менее 22%. Постоянное уточнение представлений о спектре возможных задач пользователя является важной частью развития современной поисковой системы, позволяющей наращивать долю задач пользователя, в которых использование поисковой системы может быть полезным. Как показано в данной работе, пользователи посещают страницы различного типа поразному. Понимание типичных путей, которыми пользователи достигают целевых страниц, может помочь также и в понимании финальных целей пользователей, задавших поисковый запрос.

В данной статье производится анализ пользовательского навигационного поведения в сети Интернет и предлагается новый метод анализа пользовательского трафика, развивающий стандартный метод, основанный на анализе источников переходов. Каждый визит страницы имеет не более одного источника, содержащего адрес предыдущей посещенной страницы (страницы, с которой был совершен переход на данную по ссылке). Значение источника может отсутствовать (например, в случаях прямого ввода адреса страницы в браузер, открытия страницы

из закладок и т.п.). Рассматривались несколько типов источников:

- *social* – страница-источник расположена на домене одной из популярных социальных сетей;
- *search* – страница-источник принадлежит поисковой системе;
- *internal* – страница-источник расположена на том же домене второго уровня, что и страница-получатель;
- *external* – страница-источник расположена на другом домене второго уровня;
- в отдельный тип *none* выделялись случаи отсутствия источника.

Как показано в данной работе, множество значений источников визитов, входящих на станицу, достаточно хорошо характеризует роль данной страницы в сети Интернет. Однако ещё больше информации о роли страницы можно извлечь, анализируя не только непосредственные источники переходов, но и полные маршруты пользователей от начала сеанса работы до достижения целевой страницы. Для этого в данной работе предлагается использовать обобщенные маршруты, позволяющие компактно хранить и анализировать ключевую информацию обо всём пути пользователя – с акцентами на точке начала маршрута, точке достижения домена (сайта) целевой страницы и непосредственном предшественнике (источнике) целевой страницы. Показывается, что различные распределения частот обобщенных маршрутов характерны для страниц с различным содержимым и функциональностью. Более глубокое понимание этих связей может позволить расширить множество задач пользователей, поддерживаемое функциями поисковых систем.

Проведенное полномасштабное исследование большого числа различных страниц сети Интернет с точки зрения навигационных паттернов их посещения показывает, что все множество страниц можно рассматривать как совокупность нескольких существенно различных кластеров страниц разного типа: новости, форумы, социальные сети, контент для взрослых, сервисы поисковых систем и т.п. Интересно, что обнаруживается

также крупный кластер, в посещении страниц которого в настоящее время никак не участвуют поисковые системы.

Среди основных результатов данной работы можно перечислить следующие:

- Предложен формализм обобщённых маршрутов для анализа пользовательского поведения в сети Интернет.
- Показано, что на основе распределения входящих визитов по типам обобщенных маршрутов страницы сети Интернет могут быть автоматически сгруппированы в несколько кластеров. Страницы в каждом из этих кластеров играют определённую роль в сети Интернет.
- Проведен анализ найденных кластеров с точки зрения навигационных сценариев и достижимости страниц кластера с помощью поисковых систем. В частности, обнаружен крупный кластер, страницы из которого почти никогда не появляются в результатах поиска. Углубленное изучение этого кластера может помочь в дальнейшем улучшении сервисов, предоставляемых поисковыми системами.

Дальнейший материал организован следующим образом: в следующем разделе описаны предшествующие исследования пользовательского поведения в сети Интернет. В третьем разделе описаны использованные в данном исследовании данные, в четвертом разделе изложен предлагаемый подход к описанию и классификации навигационных маршрутов.

Основная часть исследования представлена в разделах 5 и 6: сначала описываются предварительные эксперименты, а затем полномасштабная кластеризация множества страниц российского сегмента сети Интернет. В седьмом разделе приводятся выводы и обсуждаются возможные направления дальнейших исследований.

2. Обзор существующих работ

Навигационное поведение пользователей сети Интернет и влияние поисковых систем на это поведение является темой ряда различных работ.

В [12] показано, что поисковые системы влияют на 13,6% пользовательских переходов. Сюда включены посещения страниц поисковых систем, навигация по страницам с результатами и посещение результатов с помощью перехода по ссылкам со страницы с результатами поиска. Показано также, что одна пятая всех сайтов, посещаемых пользователями, посещается только через поисковые системы. Приводятся различные характеристики навигационных сессий, в том числе распределение длины сессий, среднее время на каждой странице в сессии, число страниц в сессии и число уникальных сайтов в сессии.

Авторы [4] показывают, что широкая популярность поисковых систем смещает общую активность (посещаемость) в сети в сторону популярных сайтов. Они оценивают влияние этого фактора на эволюцию веб-страниц и приходят к выводу о том, что в случае ранжирования результатов поиска по популярности новой странице требуется значительное время для получения трафика, даже если качество страницы очень велико. Исходя из этого, можно сделать вывод, что поисковые системы способны влиять на пользовательское поведение и привычки путем изменения алгоритмов ранжирования и введения новой функциональности, увеличивающей число задач, которые можно решать с помощью поисковой системы. Как следствие, могут существовать пути увеличения множества задач, в решении которых поисковые системы являются основным инструментом.

С другой стороны, в работе [1] приводятся аргументы к тому, что широкое распространение поисковых систем может создавать и уравнивающий эффект. Поисковые системы повышают шансы новой страницы быть обнаруженной пользователями в случае, если она содержит уникальный контент, релевантный интересам пользователя, выраженным в его поисковом запросе.

Согласно [7], все страницы могут быть разделены на несколько классов:

- контентный (новости, порталы, игры, вертикальные сервисы, мультимедиа);
- коммуникационный (почта, социальные сети, форумы, блоги, чаты);

- поисковый (веб-поиск, объектный поиск, мультимедийный поиск).

Авторы рассматривают различные типы значений источника перехода с целью анализа путей, которыми пользователи перемещаются между страницами внутри и между доменов и классов страниц, а также возможного воздействия поисковых систем на эти пути.

В нашей работе мы развиваем анализ, проведенный в [7], в нескольких направлениях. Во-первых, наша цель состоит в изучении навигационных профилей отдельных страниц (в [7] изучается статистика, агрегированная по типам источников). Во-вторых, вдобавок к типам источников мы рассматриваем полные навигационные пути и их классы. Наконец, мы выявляем несколько типичных кластеров страниц, сформированных по принципу подобия паттернов входящего трафика.

В [5] навигационные пути рассматриваются с пользовательской точки зрения; авторы анализируют и сравнивают поведение различных демографических групп. В нашей работе мы, напротив, концентрируемся на свойствах страниц и маршрутов, приводящих к ним. Пользовательские навигационные привычки также исследуются в [6]. Показано, что пользователи склонны предпочитать некоторые домены и могут переходить на результаты поиска с этих доменов, даже если на странице результатов присутствуют более релевантные результаты из других источников.

Навигационные пути и их значение при исследовании поискового поведения также анализируются в [2, 3, 16, 18]. В частности, в [18] с помощью различных метрик демонстрируется, что в среднем всякая страница, фигурирующая в навигационном пути, начинающемся с поиска, является важной для индексирования поисковой системой. В [3] статистика навигационных путей, начинающихся с поиска, используется для улучшения ранжирования документов на странице результатов поиска.

Статистика пользовательского навигационного поведения успешно используется для построения новых моделей релевантности и авторитетности [10]. Таким образом, анализ взаи-

мосвязей между пользовательским навигационным поведением и свойствами страниц является важным инструментом для получения и уточнения информации о роли и качестве данных страниц. Так, например, в [20] демонстрируется, как можно улучшить качество алгоритма BrowseRank, используя статистику положения страниц в навигационных путях пользователей.

В [9, 12] показано, что популярность страницы и динамика её популярности существенно зависят от расположения страницы и могут быть предсказаны на её основании. Исходя из этого, можно предположить, что и поведенческие сценарии, приводящие пользователей к странице, также зависят от типа и расположения данной страницы. Таким образом, изучение пользовательских навигационных сценариев может помочь в понимании структуры сети Интернет и роли различных страниц в ней, и, наоборот, можно пытаться предсказывать подходящие навигационные сценарии для новых или ещё не существующих страниц на основании их расположения.

Авторы [17] анализируют корреляцию между субъектом поиска, объектом поиска и поисковым поведением. Они проводят кластеризацию пользователей поисковых систем на основании поисковых запросов и анализируют результирующие кластера. Анализ, предлагаемый нами, близок по подходу, но вместо исследования множества пользователей мы сосредотачиваемся на исследовании страниц.

3. Использованные данные

3.1. ИСХОДНЫЙ МАССИВ ДАННЫХ

В работе использовались данные о активности пользователей Яндекс.Браузера, добровольно согласившихся предоставлять анонимизированную статистику, а именно, о всех посещениях ими страниц интернета за период в три месяца: с 1 августа 2013 г. по 31 октября 2013 г. Популярность различных страниц существенно различна; только 2% всех этих страниц были посещены не менее чем 20 различными пользователями за наблюдаемый период, при этом визиты на эти страницы составляют 65% всех визитов на все возможные страницы за период. Стра-

ницы с менее чем 20 различными посетителями за период были удалены из дальнейшего рассмотрения с тем, чтобы исключить из анализа технические (одноразовые) страницы, а также страницы, содержащие персональные данные (например, страницы личной почты, персональных настроек и т.п.).

Далее из множества страниц были выделены два существенно важных подмножества: поисковые страницы (страницы поисковых систем) и социальные страницы (страницы социальных сетей). Поисковые страницы важны в рамках данного исследования, так как одна из его целей – выяснение влияния поисковых систем на навигационное поведение пользователей. Социальные сети также заметно обуславливают навигацию современного пользователя сети Интернет, а сами пользователи, находясь на страницах социальных сетей, демонстрируют специфическое поведение [8].

Поисковыми считались страницы, принадлежащие домену одной из трех популярнейших поисковых систем в России: yandex.ru, google.ru и mail.ru (вместе данные системы покрывают около 97% всего российского поискового трафика). Страница была отнесена к социальным в случае, если она принадлежит домену одной из 22 наиболее популярных в России социальных сетей.

Таким образом, в дальнейшей работе использовалась информация о посещениях пользователями нашей выборки множества выбранных нами страниц. Каждая запись о посещении содержит четыре поля:

- уникальный обезличенный идентификатор пользователя;
- идентификатор момента времени визита;
- адрес посещённой страницы;
- адрес страницы-источника, с которой был осуществлён переход (поле может быть пустым в случаях, когда пользователь переходит по закладкам или вручную указывает адрес страницы в адресной строке).

3.2. ПОСТРОЕНИЕ СЕССИЙ И РЕФЕРАЛЬНЫХ ДЕРЕВЬЕВ

По аналогии с подходом, использованным в [7], будем анализировать пользовательские маршруты по рассматриваемым страницам, построенные на базе пользовательских сессий (см. рис. 1 в качестве примера).

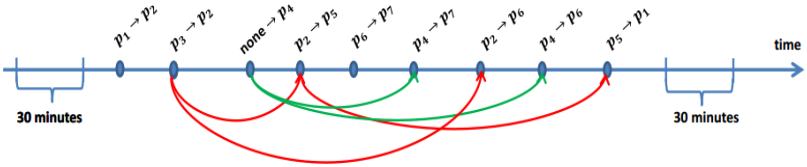

Рис. 1. Пример навигационной сессии и ее реферального леса

Определим пользовательскую сессию как последовательность посещений страниц одним пользователем, в которой пауза между двумя соседними посещениями не превышает 30 минут. Это пороговое значение – одно из наиболее широко используемых при решении задачи построения поисковых или навигационных сессий [15].

Для каждой такой пользовательской сессии определим реферальный лес (совокупность реферальных деревьев) следующим способом:

- Рассмотрим граф над множеством вершин, каждая из которых соответствует одному из посещений пользователя в рамках данной сессии и задана триплетом **<t, ref, targ>** с компонентами **t** (временная метка), **ref** (страница-источник перехода), **targ** (страница-цель перехода).
- Ориентированное ребро из **<t1, ref1, targ1>** в **<t2, ref2, targ2>** создается в том, и только в том случае, когда выполняются следующие правила:
  - **targ1 = ref2**;
  - **t1 < t2**;
  - в данной сессии не существует записи **<t3, ref3, targ3>** такой, что **targ3 = ref2** и **t1 < t3 < t2**.
- Руководствуясь этими правилами, рассмотрим последовательно все возможные пары посещений в текущей сессии и определим множество рёбер нашего графа.

В общем случае полученный граф будет лесом, так как отсутствие циклов гарантируется требованием упорядочивания посещений по времени.

Далее будем рассматривать связные компоненты этого графа (реферальные деревья) независимо.

Каждое из реферальных деревьев имеет выделенную корневую запись – это запись, в которую нет входящих ребёр, а её временная метка минимальна среди всех вершин данного дерева.

Определим тип каждого дерева по его корневой записи:

- Будем считать дерево начавшимся с поиска (search), если метка корневой записи дерева имеет вид **<*, ref = SERP, *>** или **<*, *, targ = SERP>**, где **SERP** – страница результатов поиска любой из поисковых систем.
- Для прочих деревьев началом считается timeout, если корневая запись в дереве имеет источник, отличный от пустого,
- и none в противном случае.

Наконец, определим навигационный маршрут для каждого посещения страницы (т.е. для каждого узла рассматриваемого графа) как путь от корня реферального дерева, относящегося к этому посещению, до самого этого посещения.

4. Обобщённые навигационные маршруты

Описанное в конце предыдущего раздела понятие навигационного маршрута, соответствующего конкретному посещению страницы, позволяет описать всю информацию о перемещениях с начала его сессии пользователя, приведших его к данной странице. Однако навигационные маршруты могут иметь произвольную длину и содержание, поэтому они плохо поддаются сравнению и классификации.

В настоящем разделе мы предлагаем понятие обобщённого навигационного маршрута, лишённое этих недостатков.

Определим обобщённый навигационный маршрут как совокупность трёх характеристик, рассчитываемых на основе навигационного маршрута:

- тип источника страницы (referrer);

- тип источника домена (domain referrer) – тип источника посещения первой страницы в маршруте, чей домен совпадает с доменом целевой страницы (т.е. тип страницы, с которой пользователь попал на домен целевой страницы);
- тип начала пути (тип корня реферального дерева) – search, none или timeout.

Типы начала пути были описаны в предыдущем разделе, а здесь мы формально определим типы referrer и domain referrer.

По аналогии с [7], мы будем рассматривать 7 типов значений источника целевой страницы: RefNone, RefSocialInternal, RefSocialExternal, RefSearch, RefMainPage, RefInternal и RefExternal. Описание этих типов может быть найдено в таблице 1 (см. также алгоритм 1 для определения типа referrer). Сначала осуществляется проверка, располагается ли страница-источник на одном домене второго уровня со страницей-получателем (Internal) или на разных (External). В первом случае дополнительно проверяется, была ли страница-источник главной страницей сайта (MainPage).

Далее, как было указано ранее, помечаются страницы-источники, принадлежащие поисковым системам и социальным сетям. Помимо информации о непосредственном источнике визита важно включить в рассмотрение более общую информацию о перемещениях пользователя, поэтому мы расширяем [7] путем добавления таких характеристик маршрута, как тип domain referrer (тип последней страницы, предшествующей попаданию на домен целевой страницы): DomNone, DomTimeout, DomSocial, DomSearch, DomOther; и тип начала пути: OriginNone, OriginTimeout, OriginSearch (см. описания этих типов в таблице 1).

Таким образом, каждый маршрут может соответствовать одному из 3*5*7 = 105 обобщенных маршрутов. В следующем разделе мы покажем, что частоты реализации этих обобщенных маршрутов по отношению к конкретной странице можно использовать для определения роли этой страницы в сети Интернет.

Таблица 1. Значения характеристик обобщённого маршрута

Тип источника страницы	**Описание**
RefNone	Пустое значение
RefSocialInternal	Источник принадлежит тому же домену, и тот является социальной сетью
RefSocialExternal	Источник принадлежит другому домену, и тот является социальной сетью
RefSearch	Источник является страницей результатов поиска
RefMainPage	Источник является главной страницей того же домена
RefInternal	Источник является другой страницей того же домена
RefExternal	Источник является некоторой страницей другого домена (не поиском и не страницей социальной сети)
Тип источника домена	**Описание**
DomNone	Пустое значение
DomTimeout	Разрыв сессии произошёл по превышению паузы в 30 минут
DomSocial	Переход на целевой домен произошёл со страницы социальной сети
DomSearch	Переход на целевой домен произошёл со страницы результатов поиска
DomOther	Переход на целевой домен произошёл с другой страницы (не поиска и не социальной сети)
Тип начала пути	**Описание**
OriginNone	Пустое значение
OriginTimeout	Разрыв сессии произошёл по превышению паузы в 30 минут
OriginSearch	Началом сессии является поиск

Алгоритм 1. Определение типа источника

```
if referrer = none then
    type = RefNone;
else
    if referrer ∈ a social network then
        if referrer ∈ the same domain as the target
        page then
            type = RefSocialInternal;
        else
            type = RefSocialExternal
    else
        if referrer ∈ search then
            type = RefSearch;
        else
            if referrer ∈ the same domain as the
            target page then
                if referrer = main page of the domain
                then
                    type = RefMainPage;
                else
                    type = RefInternal;
            else
                type = RefExternal;
```

На рис. 2 представлены примеры нескольких обобщенных маршрутов.

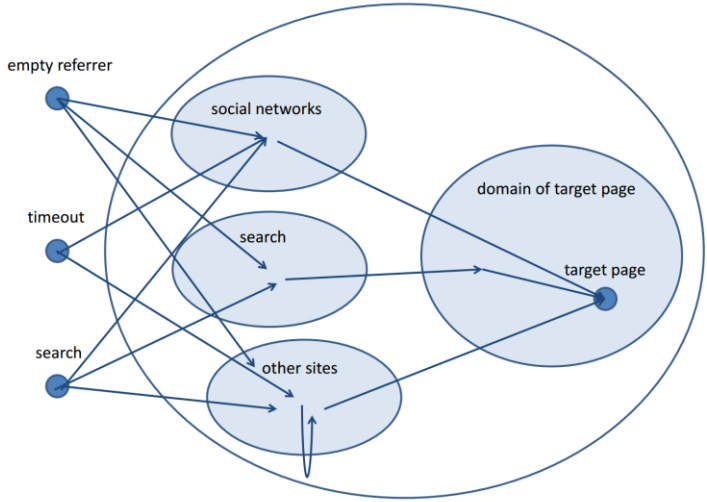

Рис. 2. Схема типов обобщённых маршрутов

Анализ экспериментальных данных показывает, что наиболее популярными обобщенными маршрутами являются:

1. (OriginTimeout, DomTimeout, RefInternal) – 29%.
   После длительного бездействия пользователь переходит по нескольким страницам в рамках одного домена.
2. (OriginNone, DomNone, RefInternal) – 19%.
   После прямого захода (через ручной ввод адреса или с помощью закладок) на домен пользователь перемещается по нескольким страницам на этом домене.
3. (OriginNone, DomNone, RefNone) – 7%.
   Пользователь попадает на целевую страницу за один шаг без referrer (через ручной ввод адреса или с помощью закладок).
4. (OriginNone, DomNone, RefSocialInternal) – 6%.
   Первый переход пользователя не содержит referrer, затем он перемещается по нескольким страницам, все из которых (включая целевую) расположены в социальной сети.

5. (OriginSearch, DomSearch, RefInternal) – 5%.
   Маршрут пользователя начинается на поиске, затем пользователь переходит на целевой домен, а затем перемещается в пределах домена до целевой страницы.
6. (OriginTimeout, DomTimeout, RefSocialInternal) – 5%.
   После длительного бездействия пользователь перемещается по нескольким страницам, все из которых (включая целевую) расположены в социальной сети.
7. (OriginTimeout, DomOther, RefExternal) – 4%.
   После длительного бездействия пользователь посещает одну или несколько страниц на одном или нескольких внешних доменах, и затем в один переход перемещается на целевую страницу на целевом домене.
8. (OriginNone, DomNone, RefMainPage) – 3%.
   Пользователь напрямую (через ручной ввод адреса или с помощью закладок) попадает на главную страницу целевого домена, а оттуда за один переход переходит на целевую страницу.
9. (OriginTimeout, DomTimeout, RefMainPage) – 3%.
   После длительного бездействия пользователь переходит на главную страницу целевого домена, а оттуда за один переход переходит на целевую страницу.
10. (OriginSearch, DomSearch, RefSearch) – 3%.
    Пользователь переходит на целевую страницу непосредственно со страницы результатов поиска поисковой системы.

5. Эксперименты

В данном разделе представлены некоторые экспериментальные результаты, подтверждающие важность и информативность выделенных нами характеристик обобщённых маршрутов (типы источников страницы, типы источников домена и типы начала пути) для описания паттернов навигации пользователей по сети и решаемых ими задач.

В частности, мы исследовали связь объективных свойств страницы с характеристиками наиболее популярных путей, при-

водящих к ней. Для этих целей рассматривались следующие базовые объективные характеристики страницы:

- Глубина вложенности страницы, рассчитанная как число символов «/» в ее адресе после домена (0 в случае главной страницы). Эта показатель, обычно называемый «глубиной адреса» (url depth), часто используется для аппроксимации реальной удалённости страницы от главной страницы сайта [19]. Мы предполагали, что этот показатель должен коррелировать с навигационной доступностью страницы, а она, в свою очередь, – отчасти определять навигационное поведение.
- Текущий возраст страницы, т.е. разность между последним днем анализируемого периода (31 октября 2013 года) и днем, когда адрес страницы был впервые обнаружен поисковой системой Яндекс (Яндекс интенсивно индексирует и переиндексирует сотни миллионов страниц ежедневно). Мы планировали проверить, влияет ли возраст страницы на то, какими способами пользователи на нее приходят.
- Размер домена, т.е. число документов, известных поиску, принадлежащих тому же домену, что и рассматриваемая страница. Эта характеристика определяет абсолютный размер домена, на котором находится страница, что, по нашему мнению, также может определять пользовательское навигационное поведение.
- Популярность страницы, оцениваемая как число различных пользователей (из нашей выборки), посетивших рассматриваемую страницу в течение анализируемого периода времени хотя бы 1 раз.

Результат анализа корреляции между перечисленными характеристиками страницы и характеристиками обобщённых маршрутов приведён на рис. 3. Здесь мы разделили значения каждой характеристики на 10 групп так, чтобы доля страниц в каждой корзине равнялась 10%. (Единственным исключением стала характеристика «глубина вложенности страницы», для которой использовалась шкала 0, 1, 2, 3, 4, 5+, так что последняя корзина содержит менее 10% страниц.) Затем были рассчитаны

доли каждого значения источников переходов на страницу в каждой корзине.

В качестве выводов анализа этих долей можно заключить:

- Страницы, более близкие к главной странице сайта, принимают большую долю внешнего трафика и меньшую долю внутреннего трафика: логично, что более глубоко размещенные страницы чаще достигаются путем внутренней навигации по сайту.
- Прямые заходы (с пустым значением источника) чаще всего происходят на главные страницы сайтов (или на ближайшие к ним страницы). Главные страницы также имеют большую долю поискового трафика, чем более глубокие страницы.
- Популярные страницы имеют высокую долю переходов с главной страницы (что логично, так как на главной странице обычно размещают ссылки на наиболее важные страницы сайта). Интересно, что менее популярные страницы имеют большую долю поискового трафика, возможно, по той причине, что их сложно обнаружить другим способом.
- Страницы с более крупных доменов получают больше внутреннего и меньше внешнего трафика. Доля переходов с главной страницы больше для страниц с небольших доменов (так как число ссылок, которое можно эффективно разместить на главной странице, ограничено).
- Более старые страницы получают большую долю внешнего трафика (старые страницы со временем становятся лучше известны за пределами собственного сайта). Старые страницы также получают большую долю визитов с пустыми источниками (т.е. через закладки или с прямым набором адреса вручную). Для молодых страниц более свойственны переходы с главной страницы сайта (например, ссылки на новости с главной страницы новостного сайта).

Далее, полностью аналогично рассматривалась взаимосвязь характеристик страниц с типами источников домена и типами начала пути. Корреляции между этими показателями можно изучить на рис. 4 и рис. 5.

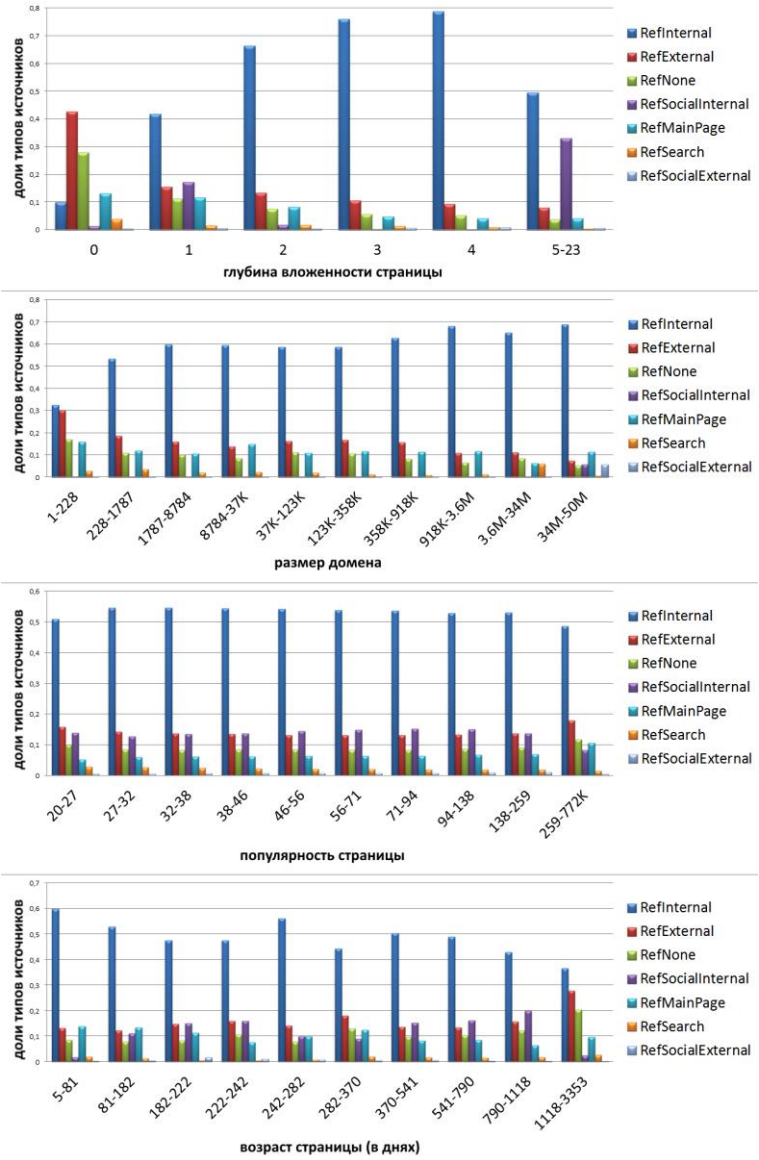

Рис. 3. Связь типов источников и свойств страниц

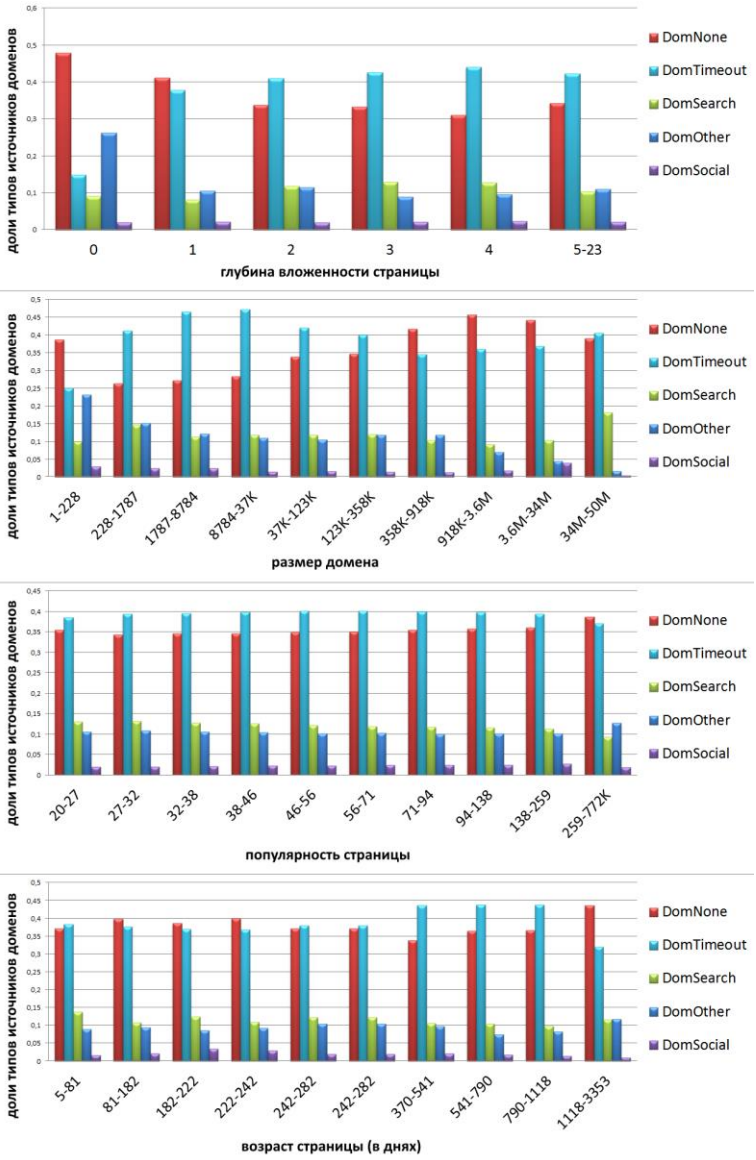

Рис. 4. Связь типов источников доменов и свойств страниц

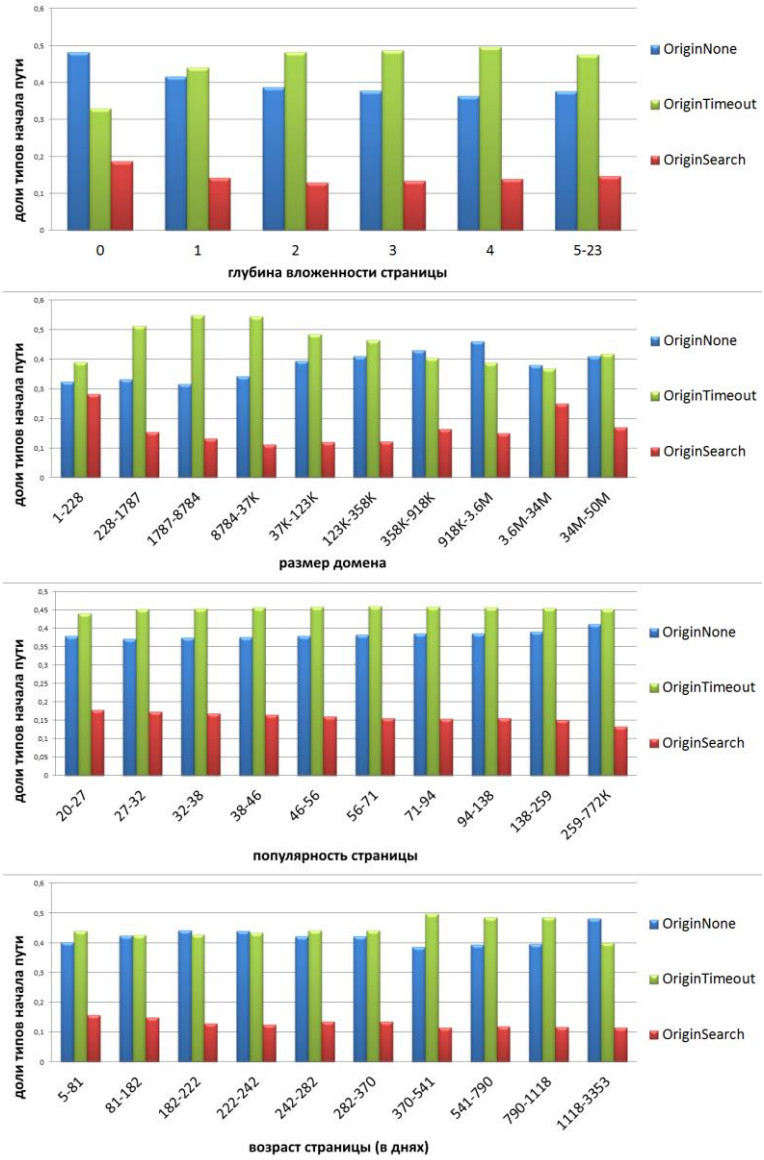

Рис. 5. Связь типов начала пути и свойств страниц

Некоторые наблюдения:

- Более старые страницы получают больший процент поисковых переходов, однако пути на молодые страницы чаще имеют поиск как начало маршрута, хотя и содержат промежуточные переходы.
- Важной представляется связь между глубиной вложенности страницы и долей поискового трафика в визитах на нее.

Как было отмечено ранее, наибольшую долю поисковых визитов имеют главные страницы сайтов, с другой стороны, наибольшая доля переходов, в которых поисковым был предварительный заход на домен, имеется у страниц с глубиной вложенности 3–4 от главной страницы. Это может соответствовать сценарию, когда пользователь достигает заглавной страницы сайта с помощью навигационного поиска, а затем перемещается к целевой странице, используя внутреннюю навигацию сайта.

6. Анализ множества страниц сети Интернет

6.1. КЛАСТЕРИЗАЦИЯ

В предыдущем разделе было продемонстрировано, что выделенные характеристики маршрутов коррелируют с базовыми характеристиками страниц. В этом разделе излагаются результаты попытки выявления групп страниц, «похожих» в терминах характеристик входящего трафика, а именно, кластеризации страниц согласно распределениям частот входящих обобщенных маршрутов.

Как было описано в разделе 3, в настоящем исследовании используется 105 обобщенных маршрутов. Таким образом, каждой странице может быть сопоставлен 105-мерный вектор длины 1, каждая координата которого равна доле входящих визитов по соответствующему обобщенному маршруту среди всех входящих визитов на эту страницу. Для дальнейших экспериментов мы сформировали из наших входных данных равномерно случайную выборку размером в 500 тысяч страниц и кластеризовали их с помощью алгоритма кластеризации expectation-maximization. Число кластеров определялось автоматически путем оптимизации с помощью перекрёстной проверки. Результа-

том кластеризации стали 7 различных кластеров. В таблице 2 представлены размеры полученных кластеров.

Таблица 2. Размеры кластеров

Кластер	**Число страниц**	**Число визитов**	**Число визитов на страницу**
1	6,8%	12,8%	311
2	4,1%	7,6%	301
3	49,6%	39,3%	130
4	10,6%	14,1%	219
5	6,2%	6,2%	164
6	21,6%	18%	137
7	1,1%	1,9%	291

С целью проверки устойчивости результатов кластеризации вся процедура была повторена для независимой случайной выборки из 500 тысяч страниц. Результатом вновь оказались 7 кластеров с сопоставимыми размерами, центрами и свойствами. Размеры кластеров двух сравниваемых разбиений и евклидовы расстояния между их центроидами приведены в таблице 3.

Таблица 3. Сопоставление кластеризаций двух выборок

	7%	**4%**	**49%**	**11%**	**6%**	**22%**	**1%**
7%	**0,08**	0,88	1,14	0,95	1,19	0,90	0,89
4%	0,93	**0,02**	0,88	0,68	0,91	0,51	0,88
48%	1,17	0,86	**0,01**	0,42	1,21	0,46	1,03
10%	0,98	0,66	0,67	**0,25**	1,05	0,46	0,81
7%	1,22	0,91	1,22	1,07	**0,00**	1,00	1,06
23%	0,96	0,53	0,43	0,30	1,01	**0,04**	0,55
1%	0,91	0,79	1,01	0,74	1,13	0,51	**0,03**

Для проверки независимости результатов кластеризации от использованного алгоритма была дополнительно предпринята кластеризация тех же выборок с помощью алгоритма *K*-means с числом кластеров, заданным равным 7. Полученные кластеры также оказались близки к результатам EM-кластеризации. Оценка согласованности Rand index [14] для между

EM-кластеризацией и *K*-means-кластеризацией для двух анализируемых независимых случайных выборок составила 0,81 в первом и 0,83 во втором случае.

6.2. ОПИСАНИЕ КЛАСТЕРОВ

В данном подразделе приводятся описания полученных кластеров, включающие в себя значения ряда их характеристик и нашу интерпретацию их специфики. Рассматривались следующие характеристики:

- типичные домены;
- типичные характеристики маршрутов, ведущих на страницы кластера;
- типичные обобщенные маршруты.

Для определения списка типичных доменов воспользуемся следующей логикой:

- Для каждого из доменов D в нашей выборке построим распределение множества его страниц по кластерам.
- Обозначим за DC долю страниц домена D, оказавшихся в кластере.
- Упорядочим для каждого кластера C все домены по убыванию значения DC.
- Будем называть наиболее типичными доменами кластера C первые *N* доменов из этого упорядоченного списка.

Понятия «типичные маршруты» и «характеристики маршрутов» вводятся по аналогии. Далее рассматриваются по 25 типичных для кластера доменов и по 5 типичных для кластера характеристик и обобщенных маршрутов.

Тематики типичных доменов кластеров представлены в таблице 4. Отметим, что имена доменов скрыты, но оставлены хорошо передающие их специфику описания. Типичные характеристики маршрутов кластеров представлены в таблице 4.

Таблица 4. Тематики типичных доменов кластеров

Кластер	Тематики типичных доменов
1	Онлайн-магазины Спортивные сайты
2	Форумы на различные темы
3	Продажа авто Мобильные сайты Коллективная закупка
4	Новостные сайты
5	Сайты для взрослых
6	Крупные социальные сети
7	Сервисы поисковых порталов

Таблица 5. Типичные характеристики маршрутов кластеров

Кластер	Типичные характеристики
1	RefMainPage, DomSearch, RefInternal, OriginTimeout, OriginSearch.
2	RefMainPage, DomSocial, RefNone, OriginTimeout, DomSearch.
3	RefInternal, OriginTimeout, OriginNone, DomSocial, DomNone.
4	RefSearch, DomSearch, RefNone, OriginSearch, RefExternal.
5	RefSearch, OriginSearch, DomSocial, RefNone, DomNone.
6	RefSocialInternal, RefExternal, DomOther, RefSearch, RefSocialExternal.
7	RefSocialExternal, DomSocial, RefSearch, OriginSearch, RefNone.

Другие свойства кластеров, в частности, средняя глубина вложенности страниц кластера, средний возраст страниц кластера и средний размер домена страниц кластера, представлены в таблице 6. В этой же таблице приводятся средние длины маршрутов, входящих на страницы кластера.

Таблица 6. Количественные характеристики кластеров

Кластер	Средняя глубина страницы	Средний возраст страницы	Размер домена	Длина пути (mean/median)
1	2,07	494 дня	3M	5,0 / 3
2	1,12	570 дней	0,5M	5,5 / 3
3	2,51	418 дней	6,7M	9,8 / 5
4	2,07	678 дней	2,3M	4,4 / 2
5	1,84	534 дня	2,2M	6,1 / 3
6	1,21	482 дня	2,9M	13,4 / 3
7	1,61	469 дней	6,7M	5,0 / 2

Далее следуют описания полученных кластеров. Каждому из кластеров присвоено условное название, отражающее специфику наиболее типичных для кластера страниц.

Например, кластер 4 получил имя «Новости», хотя это не означает, что он состоит только лишь из новостных страниц. Это означает, что в этом кластере много новостных страниц и что распределение обобщенных маршрутов у страниц из этого кластера типично для новостных страниц.

6.2.1. КЛАСТЕР 1: «ОНЛАЙН-МАГАЗИНЫ»

Данный кластер в основном состоит из страниц товаров, размещенных на сайтах интернет-магазинов. Страницы кластера чаще всего достигаются пользователем через поиск или другую форму навигации на главной странице соответствующего сайта. Сюда, например, относятся онлайн-магазины с формой поиска, а также спортивные сайты с формой для организации ставок на результаты будущих спортивных игр. Для всех страниц кластера в среднем 22% входящего трафика является переходами с главной страницы сайта, а еще 71% – переходами с других внутренних страниц того же сайта. Наиболее типичный путь к странице этого кластера выглядит так: чаще всего пользователи попадают на целевую страницу через главную страницу сайта, куда в свою очередь переходят с поисковой системы или вручную. Аналогичные выводы можно получить, анализируя наиболее типичные характеристики страниц кластера. Согласно таб-

лице 2, в данный кластер также входят в среднем достаточно популярные страницы.

6.2.2. КЛАСТЕР 2: «ФОРУМЫ»

Данный кластер в основном состоит из страниц онайн-форумов. Его страницы достаточно популярны, хотя и расположены на относительно небольших доменах. Типичный маршрут к такой странице выглядит как переход на главную страницу сайта тем или иным способом, а затем переход на целевую страницу – напрямую или через одну или несколько промежуточных страниц того же домена. Вместе с первым кластером данный покрывает почти весь трафик, проходящий через главные страницы. Из таблицы 6 также следует, что страницы данного кластера очень близки к главной странице по уровню вложенности, что достаточно реалистично для форумов. В среднем страницы этого кластера моложе, чем в других кластерах.

6.2.3. КЛАСТЕР 3: «НЕДОСТУПНЫЕ ДЛЯ ПОИСКА»

Это самый крупный кластер из получившихся. Он состоит из страниц, практически лишенных поискового трафика. В основном страницы данного кластера расположены на крупных доменах и достаточно сильно удалены от главной страницы. Пути, ведущие на эти страницы, обычно достаточно длинны и почти обязательно проходят через внутренние страницы сайта. Домены данного кластера достаточно разнообразны, так что выделить наиболее типичные затруднительно; однако среди прочих можно упомянуть крупный портал автомобильной направленности, мобильные сайты и онлайн-игры, а также сайты для организации коллективных закупок. В частности, можно отметить, что часть из этих сайтов имеют зоны, закрытые для незарегистрированных пользователей (и для индексации поисковыми системами). Так или иначе, данный кластер представляет наибольший интерес для возможного развития качества существующих навигационных сервисов.

6.2.4. КЛАСТЕР 4: «НОВОСТИ»

Этот кластер состоит в основном из страниц, принадлежащих новостным сайтам. Пути, ведущие к ним, обычно коротки: их медианная длина равна двум переходам. Пользователи часто используют поисковые системы для попадания на эти страницы.

6.2.5. КЛАСТЕР 5: «САЙТЫ ДЛЯ ВЗРОСЛЫХ»

Наиболее типичными для данного кластера являются страницы «взрослых» сайтов и других сайтов развлекательной направленности. Анализ типичных маршрутов позволяет понять, что есть два основных, приблизительно одинаково популярных пути достижения этих страниц – через поисковую систему и через главную страницу.

6.2.6. КЛАСТЕР 6: «СОЦИАЛЬНЫЕ СЕТИ»

Типичные домены этого кластера принадлежат социальным сетям и файловым хостингам. Причина их попадания в один кластер заключается в том, что в современной сети Интернет «тяжелый» контент (видео, музыка, изображения) часто сохраняется на специальных файловых хостингах, а ссылки на него распространяются через социальные сети. Поэтому распределение путей, ведущих на подобные страницы с контентом, похоже на распределение путей, ведущих на страницы социальных сетей. Страницы этого кластера обычно имеют малую глубину вложенности, но при этом пути к некоторым из них могут быть очень длинными (13,4 перехода в среднем), что может объясняться спецификой поведения пользователей социальных сетей.

6.2.7. КЛАСТЕР 7: «СЕРВИСЫ ПОИСКОВЫХ СИСТЕМ»

Анализ типичных доменов показывает наличие большого количества поддоменов популярных поисковых систем, в том числе: play.google.com, translate.google.com, docs.google.com, disk.yandex.ru, go.mail.ru и пр. Типичные пути к странице этого кластера проходят через поиск или через социальные сети; во-первых, очевидно, что на эти страницы можно перейти непосредственно с поисковых систем; во-вторых, ссылки на подобные страницы достаточно часто распространяются через соци-

альные сети (например, на сайты docs.google.com или disk.yandex.ru). Страницы этого кластера обычно расположены на больших доменах, и пути, ведущие к ним, достаточно коротки.

6.3. АНАЛИЗ ПОИСКОВОГО ТРАФИКА

В настоящем разделе предлагается более детальный анализ распределения поискового трафика по выявленным кластерам. На рис. 6 (сверху) отражено, как частота появления каждой из поисковых характеристик обобщенного маршрута (OriginSearch, DomSearch, RefSearch) распределена по рассматриваемым кластерам, например: 24% переходов с характеристикой RefSearch принадлежат кластеру 4.

Изучение рис. 6 позволяет сделать следующие выводы: чаще всего пользователи используют поисковые системы для попадания на новости (Cluster 4), страницы социальных сетей (Cluster 6) и взрослый контент (Cluster 5). Интересно, что, несмотря на то, что кластер 3 содержит примерно половину страниц (и 39% всех пользовательских визитов), он покрывает лишь 1% путей с характеристикой RefSearch, т.е. пользователи почти никогда не попадают на страницы этого кластера с помощью поисковых систем.

Рис. 6 (снизу) демонстрирует, какая доля маршрутов каждого кластера имеет ту или иную поисковую характеристику (OriginSearch, DomSearch, RefSearch): так, например, более 30% путей, ведущих в кластер 7, начинаются с поисковой системы.

6.4. АНАЛИЗ РОЛИ ХАРАКТЕРИСТИК

Для оценки влияния характеристик обобщенных маршрутов на результаты кластеризации мы рассчитали условную энтропию кластеров при условии данной характеристики. А именно, для каждой характеристики мы рассмотрели две случайные переменные – cluster, принимающую 7 различных значений, и feature, принимающую значения 0 и 1. Значения условной энтропии приведены в таблице 7.

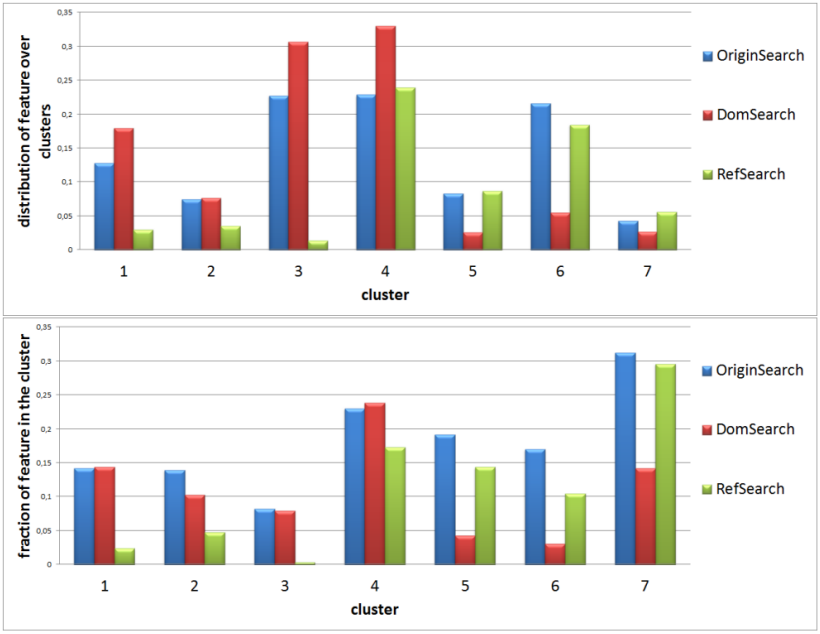

Рис. 6. Связь типов начала пути и свойств страниц

Таблица 7. Условная энтропия кластеров

Характеристика	**Условная энтропия**
RefInternal	0,606
RefSocialInternal	0,634
RefMainPage	0,684
RefSearch	0,702
RefExternal	0,708
DomSearch	0,711
OriginSearch	0,715

Чем меньше значение энтропии для характеристики, тем более информативна эта характеристика для кластеризации. Например, характеристики RefInternal и RefSocialInternal позволяют разделить кластеры 3 и 6 между собой и отделить их от других кластеров, характеристика RefMainPage важна для кластеров 1 и 2, а RefSearch – для кластеров 4 и 5. Согласно табли-

це 6 характеристики, связанные с типом referrer, наиболее важны для кластеризации – это ожидаемый результат, так как очевидно, что ближайшие соседи лучше характеризуют роль страницы в сети Интернет.

7. Заключение и выводы

В данной работе мы предложили новый метод анализа пользовательского навигационного поведения. В частности мы предложили классификацию маршрутов, ведущих к странице, акцентированную на ключевых моментах: начало пути, попадание на целевой домен, попадание на целевую страницу. На основании распределения входящих визитов по предложенным типам маршрутов, все достаточно популярные страницы российского сегмента сети Интернет были кластеризованы на несколько групп по специфике навигационного поведения пользователей по отношению к ним. Нам удалось интерпретировать полученные кластеры и сопоставить им разумные описания, связанные с типами сайтов: онлайн-магазины, форумы, социальные сети, взрослый контент, новости, сервисы поисковых порталов. Мы также обнаружили крупный кластер страниц, почти не получающих входящих визитов с поисковых систем.

Среди перспективных направлений дальнейших исследований можно отметить:

- Проведение более внимательного и детального анализа третьего кластера (страницы, недоступные для поиска) с целью выяснения, что препятствует их появлению в поисковых сценариях, какую часть из этих страниц можно включить в обслуживание поисковой системой и каким образом (например, с помощью создания новых сервисов).
- Проведение более широкого исследования, в котором каждая страница будет характеризоваться не только спектром путей, приводящих на нее, но и спектром исходящих путей – что позволит более точно определить роль каждой страницы в навигационном поведении пользователей сети Интернет.

Литература

1. BAEZA-YATES R., JR A.P., ZIVIANI N. *The evolution of web content and search engines* // Proc. of the 8th ACM Workshop on Web Mining and Web Usage Analysis. – 2006. – P. 68–73.
2. BAILEY P., WHITE R.W., LIU H., KUMARAN G. *Mining historic query trails to label long and rare search engine queries* // ACM Transactions on the Web. – 2010. – Vol. 4(4). – P. 1–27.
3. BILENKO M., WHITE R.W. *Mining the search trails of surfing crowds: identifying relevant websites from user activity* // Proc. of the 17th International Conference on World Wide Web. – 2008. – P. 51–60.
4. CHO J., ROY S. *Impact of search engines on page popularity* // Proc. of the 13th International Conference on World Wide Web. – 2004. – P. 20–29.
5. GOEL S., HOFMAN J.M., SIRER M.I. *Who does what on the web: A large-scale study of browsing behavior* // Proc. of the 6th International AAAI Conference on Weblogs and Social Media. – 2012. – P. 4–6.
6. IEONG S., MISHRA N., SADIKOV E., ZHANG L. *Domain bias in web search* // Proc. of the 5th ACM International Conference on Web Search and Data Mining. – 2012. – P. 413–422.
7. KUMAR R., TOMKINS A. *A characterization of online browsing behavior* // Proc. of the 19th International Conference on World Wide Web. – 2010. – P. 561–570.
8. LESKOVEC J., BACKSTROM L., KUMAR R., TOMKINS A. *Microscopic evolution of social networks* // Proc. of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. – 2008. – P. 462–470.
9. LIU M., CAI R., ZHANG M., ZHANG L. *User browsing behavior-driven web crawling* // Proc. of the 20th ACM International Conference on Information and Knowledge Management. – 2011. – P. 87–92.

10. LIU Y., GAO B., LIU T.-Y., ZHANG Y., MA Z., HE S., LI H. *Browserank: letting web users vote for page importance* // Proc. of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. – 2008. – P. 451–458.
11. MEISS M.R., MENCZER F., FORTUNATO S., FLAMMINI A., VESPIGNANI A. *Ranking web sites with real user traffic* // Proc. of the 9th International Conference on Web Search and Web Data Mining. – 2008. – P. 65–76.
12. OSTROUMOVA L., BOGATYY I., CHELNOKOV A., TIKHONOV A., GUSEV G. *Crawling policies based on web page popularity prediction* // In Advances in Information Retrieval, Lecture Notes in Computer Science. – 2014. – Vol. 8416. – P. 100–111.
13. QIU F., LIU Z., CHO J. *Analysis of user web traffic with a focus on search activities* // WebDB. – 2005. – P. 103–108.
14. RAND W.M. *Objective criteria for the evaluation of clustering methods* // J. of the American Statistical Association. – 1971. – Vol. 66(336). – P. 846–850.
15. SPINK A., PARK M., JANSEN B.J., PEDERSEN J. *Multitasking during web search sessions* // In Information Processing and Management. – 2006. – Vol. 42(1). – P. 264–475.
16. TOLSTIKOV A., SHAKHRAY M., GUSEV G., SERDYUKOV P. *Through-the-looking glass: utilizing rich post-search trail statistics for web search* // Proc. of the 22nd ACM International Conference on Information and Knowledge Management. – 2013. – P. 1897–1900.
17. WEBER I., JAIMES A. *Who uses web search for what: and how* // Proc. of the 4th ACM International Conference on Web Search and Data Mining. – 2011. – P. 15–24.
18. WHITE R.W., HUANG J. *Assessing the scenic route: measuring the value of search trails in web logs* // Proc. of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval. – 2010. – P. 587–594.
19. ZHU T., GREINER R., HÄUBL G. *Learning a Model of a Web User's Interests* // Proc. of the 9th International Conference on User Modeling. – 2003. – P. 65–75.

20. ZHUKOVSKIY M., KHROPOV A., GUSEV G., SERDYUKOV P. *Introducing search behavior into browsing based models of page's importance* // Proc. of the 22nd International Conference on World Wide Web Companion. – 2013. – P. 129–130.

ANALYSIS OF WEB STRUCTURE USING GENERALIZED NAVIGATIONAL ROUTES

Aleksei Tikhonov, Yandex, Moscow (altsoph@yandex-team.ru).

Abstract: *Online search engines play an increasingly pervasive role in navigating users to the web pages of their interest. Given the ambitions of any major search engine to be a "one-stop service" for all user needs, it is important to understand the ways users find content on the Web. The proposed way of generalized description of navigational patterns used to learn the ways users approach different pages on the Web depending on the characteristics of these pages. We conducted a comprehensive large-scale study of navigational profiles of different web pages and found that the Web consists of several typical non-overlapping clusters formed by pages of similar ranges of incoming traffic. These clusters can be characterized by the functionality and the purpose of their pages. This approach is useful for finding user tasks that can be supported by search systems but not currently covered by them.*

Keywords: internet analysis, user behaviour, internet navigation, web clusterization.

Статья представлена к публикации членом редакционной коллегии В. В. Мазаловым.

Поступила в редакцию 24.05.2016.
Опубликована 30.09.2016.