Поисковое поведение пользователя Яndexa (анализ веблогов)*

Николай Бузикашвили

Институт системного анализа РАН
buzik@cs.isa.ru

Изложены теоретические предпосылки, цели и результаты первого этапа анализа логов Яndexa. (1) Введены понятия логической структуры и пространства физической реализации поиска. Рассмотрены типы отображений (реализаций) логической структуры и разные типы сессий как проекции реализации на подпространства пространства реализации. Высказано предположение об ограничениях на реализацию, вытекающих из ограниченности кратковременной памяти. (2) Построена процедура автоматического выявления логических сессий и проверены гипотезы об их чередовании. Показано, что чередование с возвратом к прерванной сессии редко, а встречаясь, обычно имеет форму одноуровневой вложенности одной логической сессии в другую. (3) Кроме того, получены стандартные характеристики поискового поведения пользователя русскоязычного Веба. Дальнейшая работа: (а) построение формального языка описания поискового поведения и автоматическая трансляция в него логов; (б) автоматический анализ и выделение закономерностей в полученных описаниях.

1. Введение

1.1. Цели исследования

1. Конечная цель исследования – выявление поведенческих закономерностей по алгоритмизованному описанию поискового поведения (ПП) пользователей Яndex. Материал – логи Яndexa.
2. Производные цели. Для достижения конечной цели нужно:

- **(Ц1)** надежно выявлять в логах действия отдельного человека-пользователя, либо иметь оценку степени надежности этого;
- **(Ц2)** среди запросов одного пользователя различать запросы, различные по предмету поиска, тем самым, выделять “логические сессии”. Транзакции разных логических сессий

могут перемежаться по времени, как, например, когда пользователь параллельно ищет в одном окне одно, а в другом – другое;

- **(Ц3)** построить язык описания ПП, позволяющий представить как модификацию запросов одной логической сессии, так и временн'ую[i] (и логическую) взаимосвязь разных, в том числе ветвящихся и параллельных, сессий одного пользователя.

Цели **Ц1** и **Ц2** возникают как следствие того, что в веблогах нет информации о принадлежности запросов человеку-пользователю. В лабораторных исследованиях ПП небольших групп отслеживание транзакций производится со стороны клиента и при этом известно, кто именно сделал тот или иной запрос, т.е. проблемы соотнесения запросов и пользователей нет. При анализе поведения в Веб иногда применяется заочное выборочное анкетирование пользователей. Имея в виду, что в логе фигурирует клиент, а не пользователь, возможность и эффективность этой, к тому же выполняемой постфактум, процедуры, выглядят сомнительными.

Цель **Ц2** возникает всякий раз, когда исследователю неизвестно, что ищет пользователь. Это возможно и при лабораторных исследованиях свободного поиска, однако в них эта цель снимается применением опросников. Тем самым, **Ц2**, подобно **Ц1, —** обременительная особенность анализа веблогов. Цель **Ц3** равно актуальна и для анализа веблогов и для лабораторных исследований.

3. Сопутствующие цели. Данное исследование — первый лог-анализ пользователей поисковой машины (ПМ), ориентированной на русскоязычный сегмент Веба. Но и без того лог-исследований, привязанных к разным ПМ, в действительности совсем не много. Поэтому было бы странно не воспользоваться возможностью получить “сопоставимые” характеристики ПП пользователей Яndexa, тем самым расширив коллекцию описанного в одних терминах поведения пользователей разных ПМ.

1.2. Краткий обзор смежных результатов

Упомянем две группы работ: 1) по структуризации потока транзакций одного пользователя, а именно выделению “мультизадачности”, и 2) работы, в которых фактически дан протоязык описания последовательного (sequential) ПП.

1. “Многозадачный” поиск и выделение логических сессий. Даже в однооконном интерфейсе, работая с одним экземпляром ИПС, пользователь, в принципе, мог чередовать запросы, относящиеся к разным предметам поиска. Тем более это вероятно

при использовании многооконного интерфейса, когда, открыв пару окон, можно поочередно в одном искать одно, а в другом другое.

Под задачей понимают поиск одного предмета, т.е. логическую сессию (см. 1.3). Многозадачность – выполнение нескольких задач одним пользователем. Многозадачный поиск был обнаружен [14]: 1) при анкетировании пользователей Excite, 2) при ручной разметке запросов к Excite, 3) при поиске с посредником в онлайновом каталоге, 4) при исследовании пользователей библиотеки. Подход к многозадачности и метод ее выявления состоят в следующем: временной интервал, на котором выделяются задачи, – поисковая (в нашей терминологии, “временная”) сессия; разнесение запротоколированных транзакций по задачам выполняется вручную — либо “экспертами”, либо самими пользователями при их последующем анкетировании. При этом не делается различия между принципиально разными ситуациями, когда, в одном случае, происходит перемежающееся выполнение задач (реальная многозадачность) и когда, завершив одну задачу, пользователь приступает к выполнению другой (последовательная смена задач).

Выводы [14]: (1) многозадачный поиск носит массовый характер; (2) предметы поиска в разных выполняемых пользователем поисковых задачах могут быть схожи и несхожи; (3) сессии, в которых выполняется многозадачный поиск, длиннее однозадачных (однако как соотносится суммарная длина двух сессий – по поиску A и по поиску B – с длиной сессии, включающей поиск A, а затем B, и с длиной сессии, в которой поочередно ищутся A и B); (4) смена задач происходит в среднем 2.11 раза за сессию, (5) во многих сессиях пользователь переключается с поиска одного предмета на поиск другого *и наоборот* (выделено нами). Если в (5) имеется в виду чередование с возвратом к прерванной задаче, а не последовательная смена задач, утверждение о массовости такой манеры поиска – очень сильное. Гораздо скромнее наблюдения [13] (AltaVista, 2002 г.): из 254 произвольно выбранных сессий, содержащих *ровно два запроса*, 206 сессий были признаны при их ручной разметке многозадачными.

Членение запросов на логические сессии (задачи) производится в [13,14 и т.д.] вручную. Возможности ручной маркировки достаточно ограничены – как размером выборки, так и кругом познаний экспертов (многие запросы – “*суб-суб*культурны”).

Если наша цель – не прикидочная оценка закономерностей сочетания логических сессий, возникает необходимость автоматической разметки (ее, в свою очередь, можно выборочно контролировать вручную). В [12] предложено воспользоваться для

этой цели методами тематического атрибутирования. Т.е. вместо того, чтобы выяснить сходство двух запросов, непосредственно сравнив их между собой, предлагается сначала выполнить их тематическую категоризацию, а затем посмотреть, отнесены ли они к одной категории. Этот подход неработоспособен – разрешающая способность метода слишком низка (мало тематических категорий). Действительно, запросы любого человека, имеющего профессию, грубо говоря, распадаются на две группы – запросы, общие для всех пользователей, и запросы, связанные с профессиональной деятельностью. Последние, даже при явном несходстве должны попадать *в одну категорию*., тогда как цель состоит в том, чтобы разделить их по нескольким логическим сессиям.[ii]

В малоизвестной работе Джунга и Джо [10] сессия выделяются по семантическому сходству *самих запросов*, а не областей, к которым их можно отнести. Длины сессий в [10] оказались почти в два раза короче длин сессий, традиционно разделенных по величине интервала между соседними запросами, т.е. временная сессия содержит 1.4–1.8 логических сессий. При том, что длины временных сессий определяются тем, какой выбран порог интервала между соседними по времени транзакциями, тем не менее интересно, что приведенное в [12] значение числа логических сессий близко к числу “синтаксически” выделенных логических сессий (см. 3.3).

2. “Язык Бейтс–Фидель” [2, 3, 8] представляет собой совокупность операций модификации запроса, выполняемых при последовательном поиске.

1.3. Теоретические посылки

1. Логическая структура процесса поиска. Введем понятие логической структуры поиска как когнитивного объекта, (почти) независимого от конкретной реализации поиска, и пространства физической реализации поиска, определяемого доступной поисковой средой. *Логическая структура поиска* – это граф зависимостей, вершинами которого являются цели поиска, запросы и результаты поиска. Логическая структура представляет два противоположных типа зависимостей: зависимости “быть результатом” (изменение целей поиска как результат того, что нашла система, запрос как результат цели поиска, найденные системой документы как результат запроса) и “композиционные” зависимости, описывающие составные цели и составные запросы. Зависимости первого типа – линейные и ветвящиеся (исходящих дуг может быть несколько)[iii], второго – “сходящиеся” в том смысле, что входящих дуг больше одной.

Логическая сессия – логическая структура с относительно неизменной целью поиска.

2. Пространство физической реализации. За последние 2-3 десятилетия методы поиска изменились не сильно. Однако современный поиск радикально отличается от “традиционного”, и эти отличия есть результат изменения *пространства физических реализаций поиска*. Тогда как пространство, соответствующее “традиционному” поиску одномерно (единственная размерность – время), пространство реализаций современного – трехмерное (размерности: окна, поисковые службы, время): поиск может вестись сразу в нескольких окнах и несколькими поисковыми службами или их экземплярами (например, “парой Яndexов”).

Одна и та же логическая структура поиска допускает разные *физические реализации (отображения логической структуры в физическую)*. Можно сказать, что пространство физических реализаций было изменено 2–3 десятилетия назад силами Microsoft. По времени это совершенно случайно совпало с появлением Веб. И в результате случайного стечения обстоятельств поиск в Веб осуществляется в совершенно другом пространстве физической реализации, чем поиск в “традиционных” (созданных в 60-70е годы прошлого века) информационных системах [4].

То, что “традиционное” пространство реализации допускало только последовательный однооконный поиск с использованием одной поисковой службы, определяло не логическую структуру поиска, а то, что ее *отображения* в пространство реализаций вынужденно носило форму *линеаризации*. Современное пространство реализации допускает отображение, не нарушающее структуры ветвящегося поиска, однако, заметим, оно не позволяет отобразить без изменений “конвергентную” структуру (композиция целей и запросов). То, что доступные интерфейсы не поддерживают работу с такой структурой, не делает ее менее реальной.

Если в старом пространстве реализаций транзакция описывалась парой физических характеристик <*время, запрос*>, то сегодня — четверкой <*время, запрос, окно, поисковая_служба*>. Транзакции связаны между собой в логические сессии (не обязательно совпадающие с физическими в старом пространстве реализаций).

3. Проекции физической структуры на подпространства физической реализации: разные типы сессий и логов. Рассмотрим проекции реализации поиска в пространства физической реализации на его подпространства <время x окно>, <время x служба_поиска>. Каждую из проекций можно рассматривать как соответствующую “сессию” – сессия в данном

окне, сессия с данной поисковой службой. Обычно оконная сессия соответствует логической сессии или, в случае ветвления последней, подсессиям. В одной реализации одной логической сессии могут использоваться разные поисковые службы, например, пользователь может искать одно и то же с помощью нескольких ПМ и в (условно) локальной ИС (напр., Medline). Соответственно, сессия с данной ПМ — проекция поиска на эту ПМ, т.е. фрагмент ее лога.

4. Ограниченность памяти человека и вытекающие из этого ограничения на физическую реализацию. На рис.1 показано, как (предположительно) реализация одной и той же логической структуры поиска зависит от ресурсов памяти человека, прежде всего кратковременной памяти [1, 11].

(а) Возможное ветвление логической сессии | **(б) Реализация в современной поисковой среде** | **(в) Реализация в старой среде (линеаризация)**

Рис. 1. Пределы реализаций логической структуры поиска, налагаемые памятью человека в одно- и многооконной поисковых средах: (а) логическая структура — потенциально возможное ветвление поиска, (б) реализация в многооконной среде — ограничение по кратковременной памяти: пользователь не способен поддерживать параллельную работу со многими ветвями, (в) реализация в однооконной среде — реализуя одни ветви, пользователь забывает о существовании других

Переключение между логическими сессиями (равно как и переключение между транзакциями разных запросов) имеет мало общего с их *последовательной сменой*, при которой пользователь приступает к новой логической сессии, *только завершив* предыдущую. Ситуация точно такая же как с жонглированием несколькими шариками: едва ли так можно назвать случай, когда человек несколько раз подбросив один шарик, отложит его, потом возьмет и несколько раз подбросит следующий, после чего перейдет к другому и т.д. Метафора с жонглированием задает крайние ситуации. Переключение же между сессиями может носить (и носит) и промежуточную форму. Если при “истинном жонглирование сессиями” исключены подряд две транзакции одной из параллельных сессий, а при “псевдожонглировании” вклинивание одной сессии в другую, то чередование сессий допускает

промежуточные ситуации, например, 11221122. Автоматически возникает вопрос о степени усилий, затрачиваемых человеком при разных схемах чередования. Понятно, схема 112233 — наименее затратная, а 123123 – наиболее. Но запрос новой сессии может быть сделан как “пометка в записную книжку” – исключительно с целью не забыть ситуативно возникшую информационную потребность. И в таком случае вклинивание сессий – это способ *исключить риск* забыть, а не способ сэкономить усилия.

5. Формализация языка Бейтс-Фидель, структуризация поисковых последовательностей в терминах “букв” и “слогов”, а также модели марковского типа на “буквах” и “слогах” подробнее рассмотрены в проекте исследования Яндекса [5]. Поскольку это предмет следующего этапа лог-анализа, здесь мы на них останавливаться не будем.

2. Идея исследования

Цели и задачи настоящего исследования претерпевали изменение в процессе исследования (начиная с неформального “созерцания” исходных данных) и состоят скорее не в проверке гипотез, а в обнаружении регулярностей. Исходно имелись в виду две мало связанные между собой и прямо противоположные по претензиям на оригинальность задачи:

1. получение стандартных числовых характеристик поискового поведения пользователя поисковой машины;
2. анализ логической структуры процесса поиска.

1. Первичная цель первой задачи – получить ***унифицированные характеристики***, т.е. максимально “совместимые” и сопоставимые с характеристиками ПП для других поисковых машин (*Excite, AltaVista* и т.д.). Грубо говоря, речь идет о заполнении еще одной строки в таблице, единообразно описывающей поведение пользователей разных ПМ.

Такое стандартизованное описание осмысленно не столько как материал, сколько как *побудительный мотив* для сравнения, причем (а) не только объектов, но и процедур, в которых получены одноименные характеристики, и (б) сравнения, требующего данных, заведомо не содержащихся в унифицированных характеристиках. Различие характеристик взаимодействия с разными ПМ может быть вызвано разными причинами, например, различиями самих ПМ (язык запросов, смысл отдельных его операторов[iv]), “культурными” (языковыми) различиями пользователей. Различия разных ПМ достаточно велики, что делает не менее законным вопрос не о причинах *различия* характеристик, а о причинах их *сходства*.

Утрируя, главный источник различий и сходства – различие метода вычисления “унифицированных характеристик”. Ярчайший носитель различий — “сессия”, понятие, для протокола HTTP совершенно эфемерное (в HTTP нет сессий) и наследованное из анализа ПП в “традиционных” ИПС. Под “сессией” понимается какой-то фрагмент всех запросов пользователя (реально, клиента). Близость запросов сессии может устанавливаться либо содержательно (общий предмет поиска), либо “внешне” (близость по времени). Однако даже в последнем случае трудно сравнивать характеристики сессий в разных исследованиях, использующих для нарезки на сессии разные пороги интервалов между транзакциями.

Даже в случае “очевидных” характеристик, *на одних и тех же данных* могут быть получены несхожие оценки, и это не следствие каких-то ошибок, а результат различия методик, любая из которых сама по себе не хуже других. Поэтому вторая, более важная, на взгляд автора, цель состоит в отработке ***унифицированного инструмента*** *получения унифицированных характеристик*. (Еще один путь: собрать коллекцию веблогов, на которых ранее проводилась оценка характеристик, и выполнить перекрестный прогон — программы, обрабатывавшие логи отдельных ПМ, теперь обработали бы все логи. Очень вероятно, что на одних и тех же логах разные программы дадут различия того же порядка, что и на разных, что трактовалось как различие ПМ и/или их пользователей.)

2. Задача анализа логической структуры процесса поиска. На данном этапе наша цель состояла в автоматическом выявлении логических сессий, выполняемых одним пользователем в рамках одной временной сессии. Предмет особого интереса — ситуации параллельного выполнения логических сессий и переключения между ними. Гипотезы о пересекающихся логических сессиях:

(а) *“биологическая”*: если несколько сессий пересекаются, то одна из них “подавляет” остальные. Мотивировка и уточнение: в силу ограниченности кратковременной памяти [1, 11], человек ведет несколько параллельных сессий менее эффективно, в результате, сессии, на которые “хватает сил” оказываются излишне длинными (длиннее последовательных сессий с тем же предметом поиска и длиннее средней сессии), а сессии, на которые сил не хватает, оказываются досрочно прерванными, т.е. “короче положенного”.

(б) *“каждой сессии – по окну”*: естественно ожидать, что если несколько сессий выполняются параллельно, то каждая из них выполняется в своем окне (или окнах, если сессия ветвящаяся) и незавершенные сессии не делят одно окно друг с другом.

(в) *неравновероятное чередование* сессий: если сессии выполняются параллельно, вероятность того, что за транзакцией одной сессии последует транзакция не этой же, а другой сессии (т.е. вероятность переключения) меньше, чем вероятность продолжения текущей сессии. Более подробно формулировка задач выявления и анализа логической структуры поиска по веблогам изложена в [5].

3. Анализ веблогов

Раздел 3.2 посвящен “клон–части” исследования и содержит оценки стандартных характеристик ПП. Оригинальная часть исследования описана в разделах 3.3 и 3.4.. Изложены результаты выделения логических сессий, рассмотрены пересечение и вложенность этих сессий. В частности, показано, что (*наиболее затратная для пользователя*) схема ПП, в которой пользователь, имея открытыми несколько логических сессий, ведет сразу все, постоянно переключаясь с одной на другую, никогда не реализуется.

Временное пересечение сессий практически всегда носит форму вложенности – пользователь прерывает текущую сессию, открывает новую и к оставленной незавершенной вернется не ранее, чем завершит новую. Это *наиболее экономный* для пользователя способ ведения нескольких сессий одновременно. Еще экономнее вообще не открывать параллельных сессий. Однако когда потребность в новой сессии ситуативна и высок риск к завершению текущей забыть эту потребность, лучший способ избежать этого – открыть новую сессию *в новом окне*. Завершив ее, даже плохо помня, на чем именно была прервана первая сессия, легко, вернувшись к окну прерванной сессии, восстановить в памяти контекст поиска на том же уровне, что на каком пользователь ориентируется в непрерываемом процессе поиска (а ориентируется он довольно слабо, так, возвраты к ранее сделанным запросам отражают не то, что их результаты лучше текущих, а то, что пользователь либо вообще не помнит ранее сделанного запроса и набирает его заново, либо не помнит своей оценки его результатов).

3.1. Предварительные замечания

Предоставленные логи — выборка из логов Яndexa за период с 9 по 15 марта включительно (плюс несколько десятков полуночных транзакций 16 марта). Число транзакций этой недельной выборки сопоставимо с числом транзакций, обрабатываемых Яndexом ежедневно, т.е. выборка весьма объемна. Надеемся, что для любого клиента, включенного в выборку, в нее попали все его транзакции, и

что подавляющее большинство транзакций, относящихся к одному запросу, начинаются с транзакции с 0-й страницей результатов.

Пользователи и клиенты. Логи содержат информацию о клиентах. *Клиенты* – это имеющие доступ в интернет отдельные ПК и локальные сети (ЛС). Информации о человеке-*пользователе* логи не содержат. То, что ПК может использоваться одновременно несколькими людьми возможно, но очень маловероятно. То, что ЛС одновременно используется несколькими людьми, очень вероятно. Если в последовательности действий одного клиента имеет место пересечение по времени разных запросов, то, в случае ПК, это запросы одного пользователя (из разных окон браузера), а в случае ЛС — это, скорее всего, запросы разных пользователей.

К сожалению, ни число запросов к ПМ, ни интервалы между запросами не позволяют надежно отличить ПК, интенсивно обращающийся к поисковой машиной, от ЛС, пользователи которой делают это не очень интенсивно.

Кроме пользователя-человека возможны и пользователи-роботы. Выявление робота просто и надежно – он совершает очередную транзакцию, как только получит ответ на предыдущий запрос.

Мы браковали всех клиентов, проявляющих (хотя бы в одной из временных сессий) признаки робота. Кроме того, отбраковывались “слишком плодовитые” клиенты — как заведомо ЛС, анализировать поведение *пользователя* в которых не представляется возможным.

Адреса и порты. В полученных логах IP-порты отсутствуют, что не позволило проверить гипотезу о соотнесении логических сессий с окнами браузера и, что неприятнее, лишило нас подсказки (способа проверки) при членении на логические сессии.

3.2. Унифицированные числовые характеристики ПП

В предоставленной выборке нами было отбраковано 354 клиента-ЛС, 91 клиент-робот, 474 клиента, не сделавших ни одного запроса, а также 2 клиента, временные сессии которых содержали слишком много логических сессий (очевидно, клиента-ЛС). В результате остались 174902 клиента, которые сделали за период протоколирования 1409732 уникальных (для клиента) запроса.

Хотя в двух последних строках табл. 1 отличия Яndexa от других ПМ порядковое, мы не склонны придавать этому сколь-нибудь большое значение. Точнее, в этом отличии мы видим не отличие, а скорее тенденцию, общую для использования всех ПМ.

Табл. 1. Сравнительные характеристики поиска в нескольких ПМ

	Fireball	Excite	AltaVista	FAST*	Яndex

	1998	1999/2001*	1998	2001	2005
Объем выборки (тысяч запросов)	16253	55 1026*	993208	451	1410
Термов в запросе	1.66	2.2 2.6*	2.35	2.3	2.92 2.68**
% булевых запросов	2.55	8.54	–	–	0.88
% продвинутых запросов	25.3	9	20, в т.ч. булевы	–	2.74

* – источники сравнительных данных: [9], FAST и Excite* – из [15]
** – после нормализации, включающей “частеречевую фильтрацию”, т.е. отбрасывание предлогов, местоимений и т.п.

3.2.1. Использование синтаксических конструкций языка запросов Яndexa

Язык запросов Яndexa предоставляет булевы (*И, ИЛИ, НЕ*), “булевы” (+ и –), продвинутые (скобки, слэш и кавычки), а также суперпродвинутые (#link, #url, $title и т.д.) операторы. (Классификация наша и совершенно условная – например, оператор скобки с не меньшим успехом можно отнести к “булевым”).

3.2.1а. Булевы и “булевы” операторы. В табл. 2 доля запросов (клиентов) по всем операциям подсчитывалась как число запросов (клиентов), в(у) которых встретилась хотя бы одна операция, поэтому сумма по строке может превосходить данные в заголовке.

Табл. 2. Процентные доли “булевых” запросов (0,88%) и доли клиентов, их использующих (3,01%)

	И (&)	*ИЛИ (|)*	*НЕ (~)*	*PLUS (+)*	*MINUS (-)*
Доля запросов	0.27	0.03	0.01	0.47	0.25
Доля клиентов	1.06	0.12	0.05	1.27	0.65

Поскольку булевы операторы допускают более одной “естественной” записи, в частности, разнятся в языках запросов разных ПМ, мы проверили использование ошибочных булевых операторов, а именно термов *and, or и not*. Оказалось, что число ошибочных записей ничтожно, а именно не превысило 0,01% запросов, а употребили их 0,04% клиентов. (Запросов, содержащих *and* – 54, *or* – 13, *not* – 18, а клиентов, соответственно, 39, 11 и 19.)

Доля собственно булевых операторов в запросах к Яndexy оказалась *десятикратно меньше*, чем в логах других ПМ. Возможная причина названа в предыдущей главе. Поскольку процесс изживания/смягчения булевых проводится и в других ПМ,

вероятно, обследования их логов, выполненное сегодня, дало бы результаты, более близкие к полученным на логах Яndexa.

В то же время достаточно высока доля “булевых” операторов – “твердых унарных *И* и *НЕ*” (+ и –). Тем самым, потребность в булевой логике сохраняется, а когда операторы, “омягчаясь”, утрачивают прозрачный и устойчивый смысл, их перестают использовать. Пользователи, использующие + и – делают это почти вдвое активнее пользователей, использующих собственно булевы операторы — соотношение долей в запросах и долей среди клиентов составляет для булевых примерно 1 к 4, а для + и – почти 1 к 2–3.

3.2.1б. Продвинутые операторы. Синтаксис оператора кавычки крайне удачен: пользователь, не знающий о его назначении, употребляя эту конструкции, употребляет ее точно в том смысле, что предписывает язык запросов. Результат в табл. 3:

Табл. 3. Процентные доли продвинутых запросов (2.71%) и доли клиентов, их использующих (9.24%)

	скобки	слэш	кавычки
Доля запросов	0.01	0.00	2.70
Доля клиентов	0.04	0.01	9.21

Доля запросов, содержащих одновременно как булевы (и “булевы”), так и продвинутые операторы, составляет 0.84%, тогда как доля клиентов, одновременно использующих и то и другое ниже — 0.67%. Применительно к раздельному использованию этих операторов, доли клиентов выше долей запросов в 2–4 раза!

3.2.1.в. Суперпродвинутые операторы

Напомнив, что общее число уникальных запросов 1409732, приведем данные по числу запросов с суперпродвинутыми операторами. Всего таких запросов 415 (0.03%). Из них содержат

$title	48 (0.0%)	#keywords	0 (0.00%)	#link	321 (0.02%)
$anchor	3 (0.0%)	#abstract	0 (0.00%)	#url	39 (0.00%)
#mime	0 (0.00%)	#hint	0 (0.00%)	#image	4 (0.00%)

3.2.2. Стандартные гистограммы

Приведем стандартные для лог-анализа гистограммы распределения числа термов в запросах и распределения просматриваемых страниц. Значения в табл. 4 и 5 примерно те же, что и для любой другой ПМ.

Табл. 4. Распределение числа термов в исходных и нормализованных запросах

число термов	исходный запрос	нормализованный запрос	число термов	исходный запрос	нормализованный запрос
0	0.00	0.09	7	0.69	0.43

1	7.75	8.16	8	0.35	0.19
2	9.69	12.19	9	0.19	0.10
3	7.81	9.20	10	0.11	0.05
4	4.85	4.94	11	0.06	0.03
5	2.60	2.23	12	0.04	0.02
6	1.33	0.95	13	0.02	0.01

Табл. 5. Доля просмотров i-й страницы результатов среди просмотров всех страниц результатов

# страницы	% просмотров этой стр.	# страницы	% просмотров этой стр.	# страницы	% просмотров этой стр.
0	68.25	7	1.13	14	0.28
1	12.95	8	0.80	15	0.21
2	6.52	9	0.65	16	0.18
3	3.85	10	0.49	17	0.15
4	2.54	11	0.40	18	0.13
5	1.75	12	0.33	19 и больше	1.53
6	1.30	13	0.28		

3.2.3. Загадочная неделя

Предоставленные данные покрывают неделю с 9 по 15 марта. Предшествующие 4 дня — праздники. 12 и 13 марта – выходные.

Табл. 6. Распределение запросов и клиентов

Дата (день недели)	число транзакций	число клиентов
9 (ср.)	365657	57161
10 (чт.)	372446	58219
11 (пт.)	359906	57061
12 (сб.)	237068	36023
13 (вс.)	244396	35637
14 (пн.)	398455	60352
15 (вт.)	394184	59912

Нас интересовало не столько остаточное дыхание праздников (дни сразу после праздников чуть менее нагружены), сколько то, нельзя ли надежнее выявить “домашних” клиентов, заведомо не содержащих пересечения логических сессий разных пользователей. Доля домашних клиентов[v] и их скоррелированность в выходные должны быть выше. Элемент a_{ij} приведенной ниже матрицы – это то, какая часть работавших в день i, работала и в день j.

	9	10	11	**12**	**13**	14	15
9 марта	100.0	44.8	40.2	**19.1**	**17.2**	38.1	36.1
10 марта	44.0	100.0	44.0	**20.3**	**17.8**	39.2	37.3
11 марта	40.2	44.9	100.0	**23.1**	**18.9**	40.9	38.7

12 марта	**30.4**	**32.8**	**36.6**	**100.0**	**35.7**	**33.4**	**31.4**
13 марта	**27.7**	**29.1**	**30.2**	**36.1**	**100.0**	**35.8**	**31.8**
14 марта	36.1	37.8	38.7	**19.9**	**21.2**	100.0	44.9
15 марта	34.5	36.3	36.8	**18.9**	**18.9**	45.2	100.0

(Жирным выделены строки и столбцы, соответствующие выходным).

Предположение о большей скоррелированности внутри групп клиентов – активных в выходные и активных по рабочим дням – подтвердилось. Однако абсолютно необъяснимой для нас оказалась выявившаяся “унимодальность на рабочих днях” – чем ближе j к i, тем выше a_{ij}. Казалось бы, все элементы одной строки, соответствующие рабочим дням, должны быть примерно равны. Однако во *всех* строках элементы a_{ij} растут по мере приближения к a_{ii} и убывают по мере удаления. Вряд ли можно отнести подобную монотонность, не имеющую ни одного исключения, на счет простого совпадения. Возможно, это результат процедуры выборки.

3.3. Временные сессии

Пользователь время от времени повторяет одни и те же запросы, что естественно, когда предмет поиска динамичен, например, новости про любимого артиста. Поэтому едва ли даже идентичные запросы следует относить к одной логической сессии. Отсюда, отчасти как компромисс, мы выделяем логические сессии не на всей совокупности запросов одного клиента, а только в рамках временной сессии. Как разбить последовательность транзакций клиента на временные сессии? Простейший вариант – производить разделение, когда интервал между соседними по времени транзакциями больше некоторого порога. Но как выбрать порог? Обычно его априорно назначают равным 10–15 минутам.

Заметим, если в транзакции страница результатов отлична от нулевой – это просмотр результатов ранее сделанного запроса. (Когда для данного запроса нулевая страница встречается повторно, такую транзакцию можно интерпретировать и как повторно сделанный запрос, и как возврат к просмотру нулевой страницы.)

Запрос, в первом вхождении которого во временную сессию страница результатов отлична от нулевой, назовем *разорванным*. Разрыв — признак неверной сегментации на временные сессии (начальные и последующие просмотры результатов одного запроса оказались в разных временных сессиях). В идеале процедура сегментации не должна порождать разрывы, каким бы ни был интервал между листаниями результатов запроса. Однако редко кто не забывал закрыть браузер и завершал просмотр только на

следующий день. Как распределены промежутки времени до транзакции с ненулевой страницей от транзакции, непосредственно предшествующей ей по времени (не важно, соответствуют эти транзакции одному или разным запросам)?

Табл. 7. Гистограмма интервалов от предшествующей транзакции до транзакции с ненулевой страницей

Время (в часах) с момента предыдущей транзакции данного клиента	Число и доля транзакций с ненулевой страницей, попадающих в этот интервал	Кумулятивная доля транзакций с ненулевой страницей, попадающих в этот или предшествующие интервалы
0.0< <=0.5	823573 (97.8%)	97.82%
0.5< <=1.0	10696 (1.3%)	99.09%
1.0< <=1.5	2950 (0.4%)	99.44%
1.5< <=2.0	1231 (0.1%)	99.58%
2.0< <=2.5	684 (0.1%)	99.66%
……	……	…..……
6.0< <=6.5	47 (0.0%)	99.83%
6.5< <=7.0	50 (0.0%)	99.84%
……	……..	…….
24.0< <=24.5	27 (0.0%)	99.94%
……	…….	…….
47.5< <=48.0	13 (0.0%)	99.97%

На первый взгляд, по числу потенциально разорванных запросов, разница между 30-минутным порогом сегментации и 2.5-часовым невелика (2.12% запросов будет разорвано в первом случае и 0.34% — во втором). Однако некоторые характеристики очень чувствительны к выбору интервала сегментации. В следующих таблицах сопоставлены результаты обработки логов, полученные при двух значениях порогового интервала (30 мин и 2.5 часа).

Табл. 8. Сравнительные результаты для двух разных порогов сегментации на временные сессии

	Интервал 30 мин	Интервал 2часа 30 мин
Число выделенных временных сессий	583009	437061
Число запросов во временной сессии	2.42	3.23
Число транзакций во временной сессии	4.07	5.43
Число логических сессий	870402	802078
Число временных сессий, внутри которых есть хотя бы одно пересечение логических	32776 (5.62%)	41183 (9.42%)
Число логических сессий во временной	1.49	1.84
Число запросов в логической сессии	1.62	1.76

Число транзакций в логической сессии	2.73	2.96

Табл. 9. Гистограммы длин логических сессий для двух разных порогов сегментации на временные сессии

Длина лог. сессии, измеренная в запросах	30 мин (%)	2 часа 30мин (%)	Длина лог. сессии, измеренная в транзакциях	30 мин (%)	2 часа 30 мин (%)
1	69.07	67.23	1	53.52	51.40
2	16.07	16.32	2	17.64	17.73
3	6.55	6.85	3	9.13	9.31
4	3.22	3.48	4	5.44	5.63
5	1.78	1.99	5	3.54	3.73
6	1.06	1.22	6	2.42	2.58
7	0.67	0.80	7	1.73	1.88
8	0.44	0.53	8	1.27	1.40
9	0.30	0.38	9	0.98	1.09
10	0.20	0.26	10	0.77	0.86

Табл. 10. Средняя продолжительность (в минутах) логических сессий, измеренных в запросах и транзакциях

Длина лог. сессии, измеренная в запросах	30 мин (мин)	2 часа 30мин (мин)	Длина лог. сессии, измеренная в транзакциях*	30 мин (мин)	2 часа 30 мин (мин
1	2.2	4.5	1	0.0*	0.0
2	7.7	15.3	2	4.6	10.6
3	12.5	25.0	3	8.1	16.8
4	17.0	34.9	4	11.4	22.7
5	21.0	43.6	5	14.5	28.3
6	25.2	52.5	6	17.1	33.4
7	28.7	60.5	7	19.9	38.5
8	32.7	70.3	8	22.1	42.3
9	36.9	75.3	9	24.5	47.2
10	40.8	88.5	10	26.7	50.0

*Длительность сессии измерялась от первой и до последней транзакции сессии. Поэтому продолжительность сессии из одной транзакции нулевая.

То, что длина *логических* сессий при большем интервале оказалась не просто больше, а *двукратно* больше, требует проверки.

3.4. Логические сессии

Мы выделяем логические сессии по попарному сходству запросов. Заполнив матрицу попарного сходства запросов данной временной сессии, мы строим транзитивное замыкание (нетранзитивного) отношения сходства. Запросы, входящие в одну компоненту связности, мы относим к одной логической сессии.

3.4.1. Сходство запросов

Как определялось попарное сходство? Сначала мы делали это следующим образом. Запросы нормализовывались, при этом отбрасывались второстепенные части речи (акцент на частеречевой фильтрации). После чего для любой пары запросов мы смотрели, есть ли среди оставшихся после нормализации слов общие, и если да, считали запросы сходными. Очевидное опасение, что несхожими будут признаны синонимы, в том числе составные (из нескольких слов). Однако оказалось, что проблема не в синонимах, а в том, что пользователь делает много опечаток, часто набирает “как слышится”, доля словарных слов относительно мала, а эвристический компонент нормализатора не всегда надежен.

Чтобы обойти эту проблему в задаче определения взаимного сходства запросов, словарь не только не нужен, но вреден. (На самом деле, он конечно нужен, но как стоп-словарь для отбрасывания неинформативных частей речи.) Именно, строка запроса переводится в нижний регистр (как и исходно), из нее выбрасываются все пробелы и пунктуации, после чего для любой пары полученных строк выясняется их “скоррелированность” – наличие одинаковых подстрок. Если эти совпадения приходятся не на “вспомогательные” слова (стоп-словарь) и выше некоторого порога (по крайней мере, в одном из запросов покрывают “почти-слово”), то запросы считаются сходными. Хотя для определения сходства запросов достаточно этого, как оказалось, эффективного подхода, тем не менее, включив его, мы не стали отбрасывать и нормализацию. Только теперь она оставлена для работы на следующем, планируемом этапе, где потребуется прежде всего для эвристического частеречевого отнесения. Конкретно, она нужна, если в язык Бейтс-Фидель внести операторы смены части речи, что интересно и доступно при изучении поиска, использующего флективные языки, в т.ч. русский (см. раздел 4).

Реально же с синонимией мы столкнулись в очень специфических формах, 1) “субкультурная синонимия”: типа “гулькин хвост = Марина Королева” (в принципе, покрывается обычным tf*idf анализом терминов, но требует работы с индексом ПМ, нами не выполнялось); 2) “многоязычие”: одно и то же слово вводилось в запросах то по-русски, то по-английски, (покрывается многоязыковым словарем, хотя мы этого и не сделали); 3) использование (заведомо не включенных ни в какие словари) аббревиатур, например, в одном запросе “magic file renamer 5.02”, а в другом “mfr502”. Обнаруживать сходство таких запросов мы не умеем. Наконец, ошибки регистра (“tkmwby” вместо “ельцин”).

Вернемся к частеречевой фильтрации. В нашем подходе прилагательные не отбраковываются. В результате, у одного из клиентов в одну логическую сессию были включены запросы “*голая Собчак*” и “*голая Хакамада*”. Что, в данном случае, правильно.

Столь вольное назначение попарного сходства и последующее не менее вольное построение логической сессии целиком основаны на соображении, что если общность слов двух произвольных запросов *разных* пользователей не только *не исключает, но* никак *не означает* сходства запросов (т.е. имеет место не отношение *сходства*, а отношение *совместимости*), то когда *пользователь один и тот же, а запросы сделаны почти в одно и то же время*, такое сходство едва ли случайно и, скорее всего, вызвано сходством целей поиска.

3.4.2. Завышенная оценка числа логических сессий

Наконец, принадлежность запросов к одной логической сессии, явная из единства деятеля, времени действия и потенциальной допустимости сходства не пары, а большей группы запросов, может никак не проявляться ни в общности, ни в любом из перечисленных вариантов синонимии. Напр., во временной сессии рис. 2: общность цели запросов ясна из сравнения не пар, а именно всех запросов, а отношение между запросом “*рефераты*” и, например, запросом “*игровое поведение*” — это отношение не сходства, а совместимости.

```
user 913, t.sess. 8
  1        (0) игровое поведение
     2     (0) зоо и сравнительная психология
        3  (0) рефераты
     2     (0) зоопсихология сайты
  1        (0) рассудочная деятельность животных
  1        (0) изучение рассудочной деятельности животных
        3  (0) реферат
     2     (0) зоопсихология сайты
     2     (1) зоопсихология сайты
     2     (2) зоопсихология сайты
     4     (0) значение механизмов смещенной активности
  1        (0) игровое поведение животных
  1        (0) науки о поведении животных
```

Рис. 2. Фрагмент файла временных сессий ширины 3 (истинная ширина представленной временной сессии — 1)

Таким образом, процедура выявления попарного сходства выявляет сходные запросы *реже*, чем нужно (и вообще не выявляет совместимые запросы, что едва ли возможно без семантического анализа). В результате, *число распознанных логических сессий оказывается завышенным*, и, как следствие, *завышена и ширина временных сессий* (число пересекающихся логических сессий).

3.4.3. Пересекающиеся (вложенные) логические сессии

Нас интересовало, происходит ли в одной временной сессии одного пользователя чередование логических сессий с возвратом к прерванной (т.е. реальная многозадачность); насколько неслучайно происходит чередование — за транзакцией одной сессии равновероятно следует транзакция любой другой или же более вероятно продолжение сессии, выполняемой в данный момент.

Текущая ширина временной сессии – количество на данный момент уже начатых, но не завершенных логических сессий. Ширина временной сессии — максимальная из ее текущих ширин. В табл. 11 даны распределения временных сессий по ширине. При этом нужно иметь в виду завышенность числа логических сессий.

Табл. 11. Гистограммы ширин временных сессий (при двух разных порогов сегментации на временные сессии)

30-минутный порог отделения временных сессий		2.5-часовой порог отделения временных сессий	
ширина вр. сессии	число вр. сессий такой ширины	ширина вр. сессии	число вр. сессий такой ширины
1	550233	1	395878
2	30716	2	37208
3	1853	3	3416
4	177	4	464
5	23	5	77
6	6	6	14
всего вр. сессий	583009	всего вр. сессий	437061

Временные сессии неединичной ширины отписывались в файлы. На рис. 2 приведен пример ошибочного распознавания ширины временной сессии (опознана как 3, на самом деле 1).

При ручном контроле результатов автоматического выделения логических сессий выяснилось, что, начиная с ширины 3, доля автоматически выявленных сессий сопоставима с долей ошибок выявления логических сессий и ошибок отнесения клиентов–ЛС к числу индивидуальных пользователей. Истинная ширина практически всех временных сессий, которым при автоматическом отнесении была приписана ширина, большая 2, оказалась меньше. Среди правильно выявленных временных сессий ширины 2 большинство содержало не чередующиеся, а *вложенные* логические сессии. Основной результат анализа чередования логических сессий — то, что чередование практически всегда имеет форму вложенности.

В табл. 12 описано, чем возобновлялись прерванные логические сессии: продолжением просмотра результатов запроса, сделанного до прерывания, или новым запросом.

Табл. 12. Возобновление логических сессий после прерывания

30-минутный порог отделения временных сессий		2.5-часовой порог отделения временных сессий	
всего возобновлений, в т.ч.	48432	всего возобновлений, в т.ч.	67096
продолжение просмотра	14085	продолжение просмотра	19607
новый запрос	34347	новый запрос	47489

В табл. 13 отражен случай, когда прерванная сессия возобновлялась продолжением просмотра результатов запроса, сделанного до прерывания, а в табл. 14 — когда прерванная сессия возобновлялась новым запросом.

Табл. 13. Распределение числа транзакций прерванного запроса до и после прерывания (строки – число транзакций до, столбцы – после прерывания)

30-минутный порог						2.5-часовой порог					
	1	**2**	**3**	**4**	**5**		**1**	**2**	**3**	**4**	**5**
1	52.72	9.79	4.20	1.82	0.92	**1**	54.03	9.73	3.77	1.70	0.97
2	8.68	2.88	1.36	0.69	0.39	**2**	8.46	2.76	1.29	0.65	0.38
3	3.03	1.03	0.62	0.30	0.26	**3**	2.89	1.03	0.62	0.31	0.22
4	1.41	0.52	0.26	0.28	0.07	**4**	1.29	0.52	0.28	0.27	0.08
5	0.74	0.27	0.16	0.11	0.10	**5**	0.68	0.31	0.14	0.10	0.10

Табл. 14. Распределение числа транзакций последнего перед прерыванием запроса и первого запроса после возобновления (строки – число транзакций запроса до прерывания, столбцы – запроса после прерывания)

30-минутный порог						2.5-часовой порог					
	1	**2**	**3**	**4**	**5**		**1**	**2**	**3**	**4**	**5**
1	65.07	7.51	2.59	1.19	0.62	**1**	63.28	7.82	2.74	1.32	0.71
2	8.10	1.86	0.71	0.36	0.17	**2**	8.29	1.85	0.74	0.37	0.21
3	2.79	0.77	0.33	0.22	0.12	**3**	2.99	0.77	0.35	0.21	0.13
4	1.37	0.39	0.22	0.08	0.07	**4**	1.46	0.43	0.21	0.09	0.08
5	0.72	0.18	0.11	0.06	0.04	**5**	0.79	0.20	0.12	0.07	0.04
6	0.43	0.08	0.05	0.02	0.03	**6**	0.47	0.08	0.05	0.04	0.03

3.4.4. Проверка гипотез о чередующихся логических сессиях

Мы проверили гипотезы о чередующихся сессиях (2.2).

(а) красивая “биологическая гипотеза”, увы, провалилась – *все* участвующие в пересечении сессии устойчиво (хотя и не намного) длиннее непересекающихся (напр., табл. 15).

(б) гипотеза “каждой сессии – по окну”, вероятно, справедлива, но все же для проверки нужны логи с IP-портами,

(в) гипотеза о неравновероятном чередовании сессий подтвердилась в самой сильной форме — вложенности логических сессий.

Табл. 15. Средние длины более коротких логических сессий для интервально заданных длин более длинных (все длины – в запросах)

Длина большей	Средняя длина меньшей		Число логических сессий	
	Пересекающейся	неперсек.	пересекающейся	неперсек.
1–2	1.00	1.00	17112	67977
3–4	1.32	1.23	24393	41888
5–6	1.57	1.44	11537	10351
7–8	1.70	1.61	5726	3576
9–10	1.79	1.68	3299	1473

Вопрос: пересекающиеся логические сессии одного *клиента* – это сессии одного или разных *пользователей*? Понятно, что веблоги позволяют дать только вероятностный ответ на этот вопрос. Например, оказалось, что пересекающиеся сессии длиннее непересекающихся. Но с равным успехом можно утверждать как то, что сессии удлиняются вследствие увеличения когнитивной нагрузки на пользователя, так и то, что пересечение сессий разных пользователей произошло именно в силу того, что эти сессии оказались более длинными.

4. Выводы и дальнейшая работа

Завершенный этап исследования состоит из двух частей.

Первая – стандартные числовые характеристики поведения пользователей Яndexa, сравнимые с такими же характеристиками поведения в других ПМ. Выявлено незначительное отличие в длинах запросов и резкое отличие в использовании специальных синтаксических конструкций языка запросов. Данные по другим ПМ были получены порядка 5 лет назад, и можно предположить, что с тех пор ПП в этих ПМ претерпело сходные изменения.

Что касается характеристик временных сессий, то вместо сравнения этой эфемерной конструкции с результатами других исследований мы получили характеристики для разных интервалов сегментации на временные сессии для одних и тех же логов Яndexa.

Гораздо более полезным результатом первой части исследования является не столько оценка характеристик поискового поведениея, сколько наработанный унифицированный инструментарий.

Вторая завершенная часть исследования представляет собой попытку реализовать в операциональном виде концепцию

логической структуры поиска: построить эффективную процедуру автоматического выделения логических сессий, и располагая такой структуризацией, проверять гипотезы о том, как сессии развиваются во времени и как “взаимодействуют” (пересечение и вложение сессий). Как кажется, удалось построить достаточно эффективный инструмент выявления логических сессий.

Изучение логических сессий показало, что

- их пересечение и вложенность крайне редки (даже с учетом того, что отбраковка ЛС неточна и какая-то доля пересечений — это пересечения сессий нескольких пользователей, работавших с Яndexом одновременно с одного клиента);
- когда имеет место пересечение логических сессий, они, как правило, не чередуются, тем более не чередуются в случайном порядке, а наиболее эффективно, с точки зрения экономии ресурсов пользователя (кратковременной памяти), вкладываются одна в другую, т.е. возврат к прерванной сессии происходит только после завершения сессии, ее прервавшей. Случаев, когда уровней вложенности более одного, чрезвычайно мало. Т.е. реальная картина поискового процесса такая: если в данный момент открыта одна логическая сессия, она может быть прервана второй, но при этом (а) возврат к первой произойдет не ранее завершения второй, (б) при двух незавершенных сессиях открытие третьей практически невозможно.

О процедуре автоматического выявления логических сессий. Она очень полезна как инструмент отбраковки временных сессий ширины 1, т.е. как средство селекции временных сессий, возможно содержащих чередующиеся логические сессии. Выполнить такую отбраковку вручную на сколь-нибудь больших выборках из логов немыслимо. В результате применения автоматической процедуры (а) объем подлежащих дальнейшему рассмотрению данных сокращается более чем в 10 раз, (б) результаты автоматической разметки (рис. 2) представлены в удобной форме, позволяющей легко выполнить (выборочную) проверку правильности этой разметки. При этом контролер выполняет гораздо более простую задачу выяснениея сходства уже выделенных сессий, а не задачу выделения сессий, как это было бы в случае независимой разметки.

На следующем этапе мы еще раз вернемся к рассмотрению чередования логических сессий на материале (почти) наверняка исключающем ЛС-клиентов, а именно на выборке, сделанной по “анти-ЛС признакам” (клиенты, работавшие в выходные, небольшая — по числу запросов, транзакций и по времени — длина временной сессии, отсутствие подозрительных интервалов между запросами).

Дальнейшая работа. Основная задача следующего этапа исследования состоит прежде всего в построении и использовании формализованной версии языка Бейтс-Фидель.

1а. Построение языка. Необходимо включить в существующий протоязык конструкции для описания нескольких сессий, в частности, ветвление. Учитывая простоту определения частей речи в флективных языках, имеет смысл включить в описание как тега “часть речи”, так и оператора смены части речи для “однокоренных” слов (например, когда “Гаусс” в запросе заменяется на “гауссово”). И наконец, формализация языка и построение процедуры преобразования потока запросов в описания на этом языке.

1б. Использование. Сессии, описанные на этом языке, уже совершенно доступны для автоматического анализа поведения [5].

2. Кроме того, на второй этап отнесено подробное теоретическое рассмотрение – с позиций принципа наименьших усилий – данных о схемах чередования (переключения между) логических сессий. В частности, в мере затратности желательно отразить не только частость переключения, но и разнозатратность того, на что происходит переключение – на продолжение просмотра или на формулировку нового запроса.

5. Благодарности

Автор считает долгом выразить признательность тем, без чьей поддержки (на разных этапах и в разной форме) его результаты оказались бы еще более скромными: Иану Рузвену, Джиму (B.J.) Джансену, Владимиру Арлазарову и Сачи Арафату. Автор признателен сотрудникам Яndexa, без участия которых эта работа вообще была бы немыслима: содействие и заинтересованность Ильи Сегаловича, Юрия Зеленкова и других членов команды Яndexa не ограничивались рамками простого предоставления данных.

6. Литература

1. Миллер Дж.: Магическое число семь, плюс или минус два. *Инженерная психология*. М. (1964)
2. Bates M.: Information search tactics. *JASIST*, 30(4) (1979) 205-214
3. Bates M.: Where should the person stop and information interface start? *Information Processing & Management*, 26(5) (1990) 575-591
4. Buzikashvili N.: Information searching behavior: Between two principles. *CoLIS’05, Glasgow 2005*. LNCS 3507, Springer (2005) 79-95.
5. Buzikashvili N.: The Yandex Study. *Workshop on Evaluating user studies in information access*, *Glasgow 2005*. Univ. of Strathclyde, BCS (2005) 48-55

6. Campbell I.: Supporting information needs by ostensive definition in an adaptive information space. *Workshop on Multimedia Information Retrieval (MIRO'95)*. BCS (1995), http://www.dcs.gla.ac.uk/~iain
7. Campbell I., Rijsbergen C.J.: The ostensive model of developing information needs. *CoLIS'02, Copenhagen, 1996*, (e.g. http://www.dcs.gla.ac.uk/~iain)
8. Fidel R.: Moves in online searching. *Online Review*, 9(1) (1985) 61-74
9. Jansen B.J., Pooch U.: A review of Web searching studies and a framework for future research. *JASIST*, 52(3) (2001) 235-246
10. Jung J.J., Jo G.-S.: Semantic outlier analysis for sessionizing Web logs. *First European Web Mining Forum, Dubrovnik 2003*. LNCS 3209, Springer (2004) (*not included*) http://km.aifb.uni-karlsruhe.de/ws/ewmf03/papers/Jung.pdf
11. Miller G.A.: Human memory and the storage of information. *IRE Transactions on Information Theory*. 3(IT-2) (1956)
12. Ozmutlu H., Cavdur F.: Application of automatic topic identification on Excite Web search engine logs. *Inf. Processing & Management* (*to appear?*)
13. Spink A., Jansen B.J., Pedersen J.: Multitasking Web search on Alta Vista. *ITCC'04*. IEEE CP. (2004)
14. Spink A., Ozmutlu H.C., Ozmutlu S.: Multitasking information seeking and searching processes. *JASIST*, 53(8) (2002) 639-652
15. Spink A., Ozmutlu H.C., Ozmutlu S., Jansen B.J: U.S. versus European Web search trends. *SIGIR Forum 2002* (2002)

The Yandex Study Technical Report: First Findings

Nick Buzikashvili

The theoretical background, goals and first findings of the Yandex TL study are described. (1) The background includes: a logical structure of search process, a space of physical search realizations, mapping of the logical structure onto the space of realizations, different types of sessions as projections of the search process in the physical search space onto subspaces of this space. Also a hypothesis of different short-term memory based restrictions of different physical realizations is formulated. (2) The term- and string-based queries similarity is introduced, and transitive closure of this relation is considered as a logical session. It is shown that switching with return to unfinished session (a) is very rare and (b) it formed as time inclusion of the second session, i.e. a searcher returns to the broken session only when he finishes a new session. (3) The Yandex study is the first TLA of Russian-language search engine and the standard quantitative characteristics of user searching behavior are

calculated. The further works: (a) design and implementation of the ‘Bates-Fidel language’ language and automatic translation of the transaction logs into this language, (b) automatic analysis of these descriptions and extraction of regularities.

---

[*] Работа подготовлена при поддержке ООО “ЯНДЕКС” (www.yandex.ru), грант 103035.

[i]Здесь и далее все “временные” в тексте читаются как “временн'ые”.

[ii] Этого достаточно, однако упомянем еще и эффект, возникающий при непрямом определении сходства, аналогичном определению близости городов по тому, на каком материке они находятся. При таком подходе Анкара окажется ближе к Сеулу, чем к Стамбулу.

[iii] С. Арафат обратил внимание автора на работы Иэйна Кэмбелла [6, 7], еще 10 лет назад на тех же основаниях (т.е. исключительно на соображении, как это должно бы быть) использовавшего ветвление для описания порождения информационных потребностей.

[iv] Пример — использование булевых операторов. (1) Зная, что *И* – это “*мягкое И*”, которое ставить не обязательно, что алгоритм его обработки мало предсказуем и изменчив, вряд ли пользователь применит эту связку. Если же он убежден, что *И* — это булево *И*, то именно в таком качестве он им и воспользуется. (2) Когда служебное слово совпадает со словом обычного языка (e.g., ‘*and’*), два априори разумных метода классификации такого слова могут дать различие в разы.

[v] Другой, тоже вероятностный, способ различения долей домашних и офисных клиентов основан на том, что, с одной стороны, большая часть клиентов не рассеяна по всем временным поясам, а находится в европейской части, а с другой, в большинстве контор жизнь пробуждается не раньше 10 утра и стихает к 8 вечера.