Санкт-Петербургский государственный университет
Математико-механический факультет
Кафедра системного программирования

Инструмент анализа пользовательских логов поисковых систем

Курсовая работа студента 445 группы
Солозобова Андрея Сергеевича

Научный руководитель
к. ф.-м. н. Грауэр Лидия Вальтеровна
Руководитель группы анализа данных
службы оценки качества поиска ООО «Яндекс»

Санкт-Петербург
2012 год

Оглавление

Введение

Поисковые системы интернета постоянно совершенствуют свои алгоритмы индексации и поиска информации в интернете. Многие из таких усовершенствования носят статистический и эвристический характер. В результате этого одной из основных проблем любой крупной поисковой системы интернета является оценка качества поиска. В конечном итоге её решение сводится к определению того, насколько пользователь доволен работой с поисковиком. На этот показатель влияет масса факторов: алгоритмы индексации и переиндексации сайтов, размер поискового индекса, алгоритмы ранжирования ответов, разнообразие предлагаемых информационных источников, свежесть предлагаемой информации, персонализация, фильтрация спама и CEO, многие другие. При внесении в них изменений в первую очередь стараются улучшить именно качество поиска, потому что это основной предоставляемый продукт. Чтобы оценить уровень и узнать сравнительные характеристики качества поиска для различных версий поисковой машины используется масса различных подходов. Строятся специальные метрики, пишутся автоматические тесты качества поисковых выдач, проводятся специальные поисковые эксперименты.

Одним из экспериментов, позволяющих сделать выводы о качестве, является детальное наблюдение за работой человека во время работы в интернете и её анализ.

Для его проведения набирается группа людей, которым даются различные задания поискового характера. После выполнения набора задач, асессора просят сформулировать, насколько ему было удобно выполнять задания. Помимо сбора отчётов и ответов по каждому заданию система, ведущая эксперимент логирует все пользовательские действия в браузере (работа с интерфейсом, клавиатурой, мышью, буфером обмена и прочее). Таким образом собирается большая подборка персонифицированной информации о том, как люди пользуются веб-поиском.

Последовательность действий, совершенных человеком в рамках работы над одним заданием называют поисковой сессией. А набор таких сессий популяцией.

После того, как эксперимент проведён, начинается непростая работа по обработке его результатов. Во многом она связана с реорганизацией записей в логах, подсчётами статистик по сессиям, их анализом и выдвижением гипотез.

В рамках данной курсовой работы, основной задачей была поставлена разработка программного инструмента, облегчающего работу аналитика при работе с логами эксперимента, позволяющего более наглядно отображать их содержимое, считать статистические данные и проводить их анализ.

1 Обзор имеющихся средств и подходов

Есть два стандартных подхода к анализу логов пользовательских сессий.

Первый заключается в последовательном детальном анализе действий совершенных пользователем в рамках отдельной сессии. В результате такой кропотливой работы могут быть обнаружены закономерности и выявлены причины того или иного поведения пользователя, на основе которых можно будет построить гипотезы и идеи оптимизаций поиска. Такой вид анализа тяжело проводить в больших масштабах. Когда же закономерность уже найдена таким способом, то сразу встаёт задача определения множества пользователей, для которых соответствующая оптимизация даст результат.

Второй подход связан с вычислением метрик и статистик на различных больших популяциях пользовательских сессий, их сравнительном анализе, построении и доказательстве статистических гипотез для полученных данных. Этот подход более масштабируем, но основная его недостаток заключаются в том, что он может нивелировать различные отклонения подпопуляций от среднего значения на объемлющей выборке. Что может привести исследователей к неверным выводам и в итоге к падению качества поиска.

В таких исследованиях всегда есть опасность принять значения, полученные на маленькой выборке данных, за глобальную тенденцию и наоборот – потерять суть зависимости или пропустить её вовсе, посчитав её значение на слишком большом наборе.

В статье [1] описывается программный инструмент, предоставляющий возможности детальной визуализации пользовательских сессий и автоматического подсчёта для них статистических показателей. Такое сочетание детального и статистического анализа сессий позволяет аналитику смотреть на данные с разных ракурсов, тем самым облегчает и делает более осмысленной его работу.

Немаловажную роль в изучении качества поиска играют позиции пользовательских кликов по элементам поисковой выдачи. На основе данных о том, в каком порядке и с какими временными промежутками человек кликает по выдаче можно судить о релевантности ответов, предпочтениях людей, их привычках при прокликивании поисковой выдачи. Эти данные позволяют существенно улучшить и разнообразить выдачу для конкретных пользователей.

2 Описание проделанной работы

В рамках данной курсовой работы написан схожий по идее с описанным в статье [1] инструмент, который позволяет проводить детальную и статистическую аналитику логов экспериментов пользовательской удовлетворённости. Так же программа умеет проводить кластеризацию сессий по кликовым статистикам и выявлять часто встречающиеся паттерны поведения пользователей.

Работу над курсовой можно условно разбить на три этапа.

На первом этапе была произведена обработка исходных данных. Так как в рамках рассматриваемого эксперимента асессоры работали удалённо и информация об их действиях присылалась на сервера специальным плагином для браузера, то логи эксперимента представляют из себя набор файлов с распределёнными в ним поисковыми сессиями различных пользователей, информацией о ходе выполнения заданий и возникающих при этом событиях. Пиково за день эксперимента информации набегало до гигабайта.

Формат таких логов избыточен и неудобен для работы. Вся информация была поделена на множество мелких файлов, по одному на каждую поисковую сессию асессора. Разбор делался в два прохода. Во время первого прохода строился индекс с информацией о том, где в каких файлах лежит информация из каких поисковых сессий. На втором проходе поочерёдно читались файлы логов и по возможности сбрасывались на диск файлы с уже разобранными сессиями. Очерёдность разбора файлов с логами определялась специальным алгоритмом, работающим на основе данных из индекса. Во время второго прохода в оперативной памяти компьютера всё время было какое-то количество данных о событиях из сессий, некоторые части которых лежали в ещё непрочитанных файлах. Целью работы алгоритма было обеспечение как можно меньшей загрузки оперативной памяти. Алгоритм работал "жадно", выбирая среди оставшихся файлов логов тот, который привнесёт меньше не сброшенных на диск данных по сравнению с остальными.

Вторым этапом было построение пополняемого индекса с необходимой информацией о сессиях (местоположение на диске, данные, статистики), позволяющей реже считывать с диска сессии и не держать их в оперативной памяти.

Был организован кэш сессий для ускорения работы в режиме детального просмотра событий, выделения общих подпоследовательностей и кластеризации.

На третьем этапе был сделан интерфейс для управления популяциями реализованы два варианта просмотра содержимого сессий в виде столбчатых диаграмм, придуманы и реализованы инструменты кластеризации сессий и подсветки самых частых последовательностей событий определённой длины.

1. Столбчатая диаграмма последовательностей событий (рис. 1). Слева-направо располагаются столбцы сессий, в которых последовательно сверху вниз идут произошедшие в рамках сессии события. Раскраска событий в цвета соответствует цветовой легенде слева. Имеется возможность исключать из отображения любые типы событий или приглушать их цвет, чтобы он не бросался в глаза.
   Любое событие в диаграмме можно выделить мышью и тогда в правой части панели отобразится детальное описание данного события.
   Так же программа умеет по задаваемой аналитиком длине находить самую частую среди всех подпоследовательность действий и подсвечивать все её встречи в рассматриваемых сессиях.

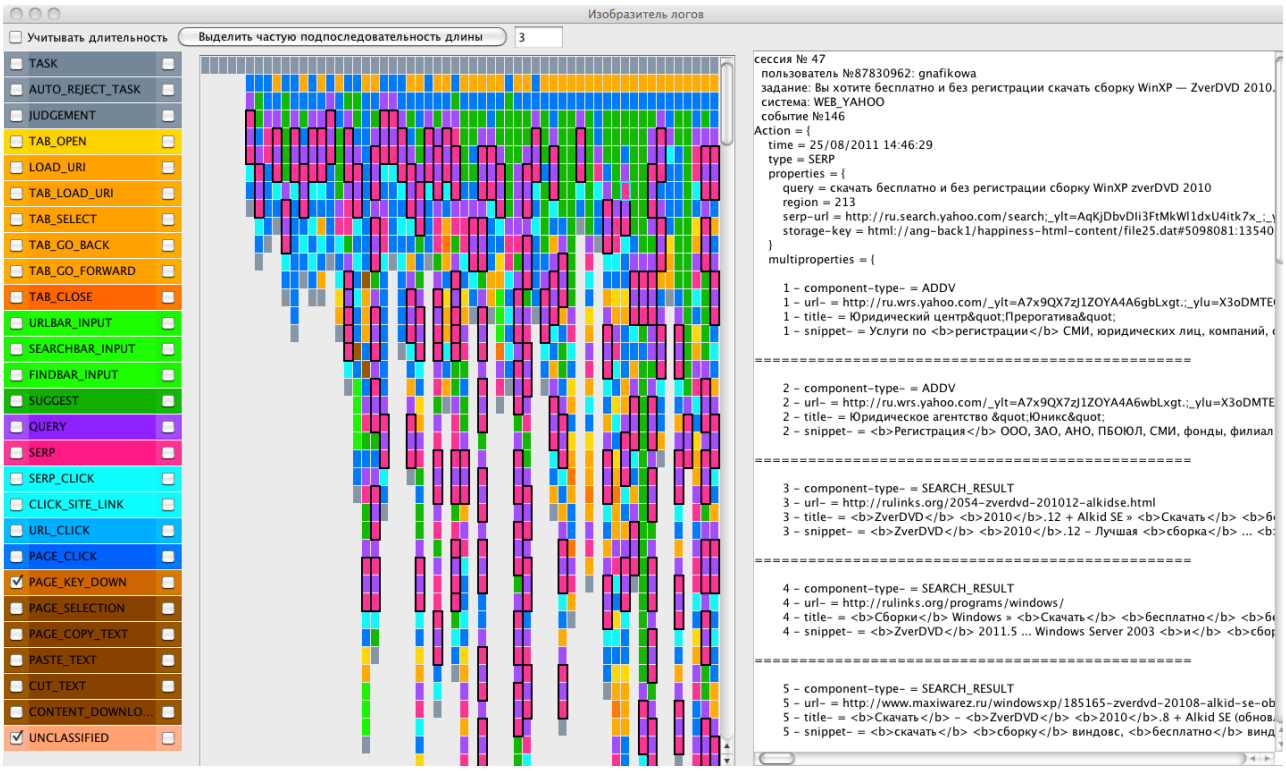

Рис. 1

2. Диаграмма хронологии событий (рис. 2). Идея и функциональность такие же, как и в предыдущем пункте, только события позиционируются согласно хронологии их происхождения и изображаются цветными рисками.

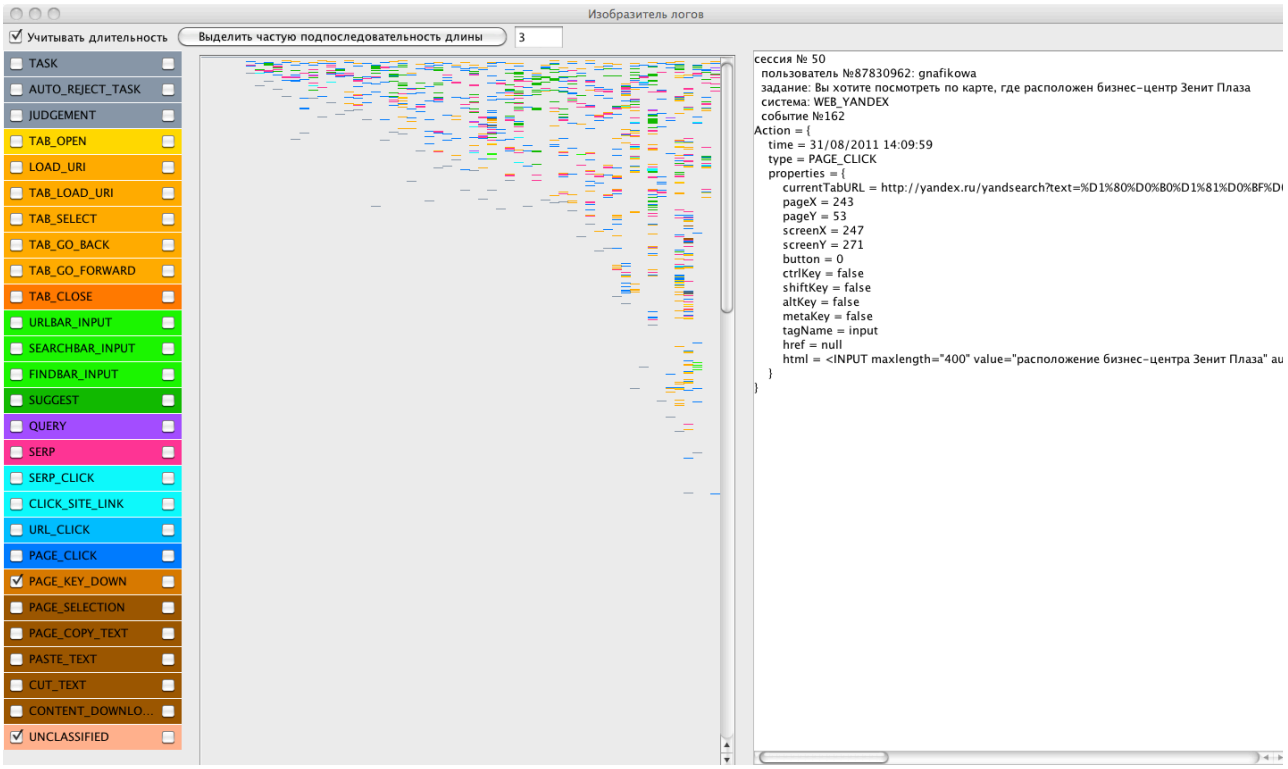

Рис. 2

3. Основная панель (рис. 3), позволяет предварительно сортировать сессии в популяциях, выбирать интересные для просмотра, сохранять их для будущих действий, выполнять различные операции над популяциями, как над множествами (объединять, вычитать, кластеризовать).

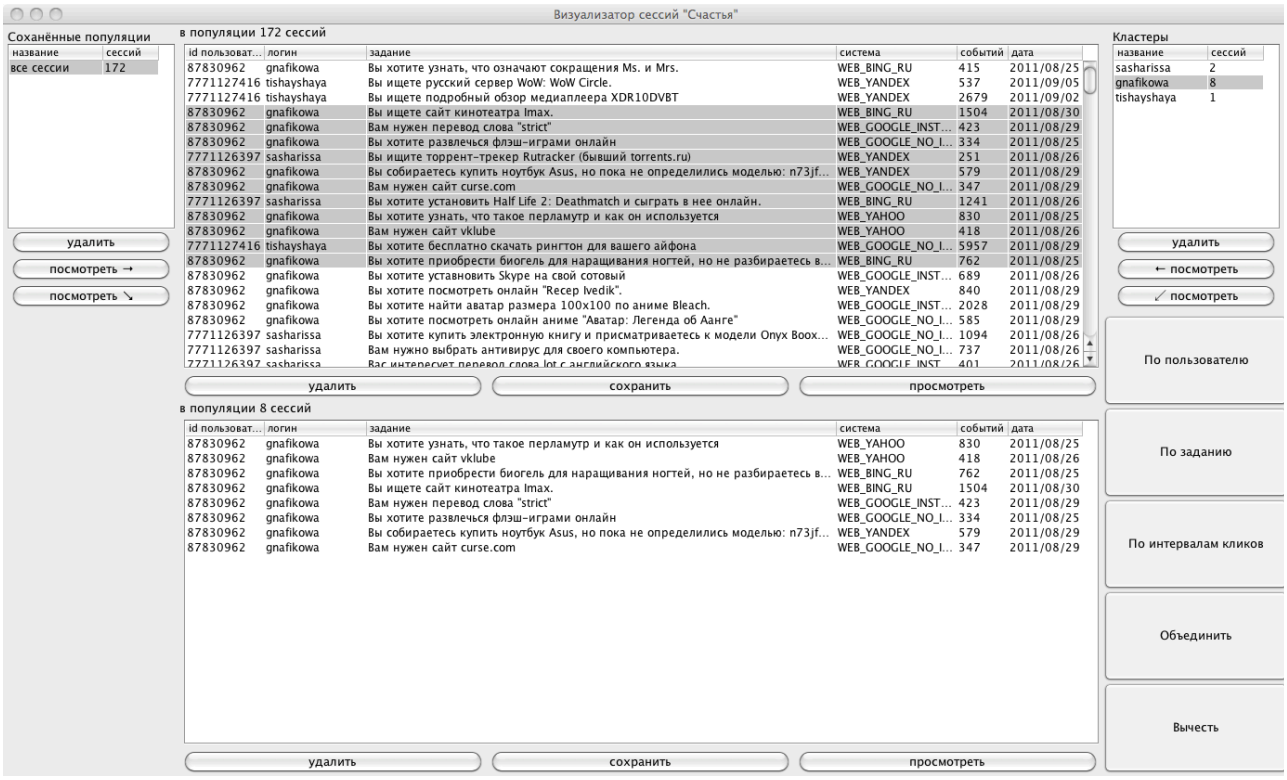

Рис. 3

4. Панель кластеризации (рис. 4) позволяет разделить выбранную популяцию на заданное количество кластеров методом К-средних. Кластеризация сессий производится в двухпараметрическом пространстве. Первый параметр – это средняя длина временного промежутка между кликами человека по элементам поисковой выдачи, второй – среднеквадратическое отклонение длин временных промежутков между кликами человека по элементам поисковой выдачи. Для сессии итоговые значения каждого показателя вычисляются как среднее арифметическое показателей всех просмотренных в ней выдач. Затем значения каждого параметра на всей популяции нормализуются вычитанием из них наименьшего и затем делением на наибольшую из полученных величин. Таким образом получается пространство двухмерных векторов со значениями из [0;1]x[0;1] .В качестве метрики на введённом пространстве используется евклидова метрика.

   Перед началом работы алгоритма, исходя из предоставляемой программой визуализации рассматриваемого множества, аналитик должен определить подходящее количество кластеров, на которое алгоритм будет делить популяцию. Начальные положения центров кластеров выбираются программой случайно. По причине рандомизированности алгоритма и того, что удачное количество кластеров не всегда удаётся подобрать с первого раза, аналитику предоставляется возможность менять количество кластеров и перезапускать кластеризацию, пока он не добьётся желаемого результата. Так же он может остановить алгоритм на любом шаге, если посчитает имеющееся разбиение удачным.

   Помимо изображения элементов множества и цветового их разбиения на кластера, на каждом кластере изображается эллипс, чей центр обозначается крестиком и показывает текущие центр кластера, а длинам полуосей соответствуют значения дисперсии элементов кластера в соответствующем направлении.

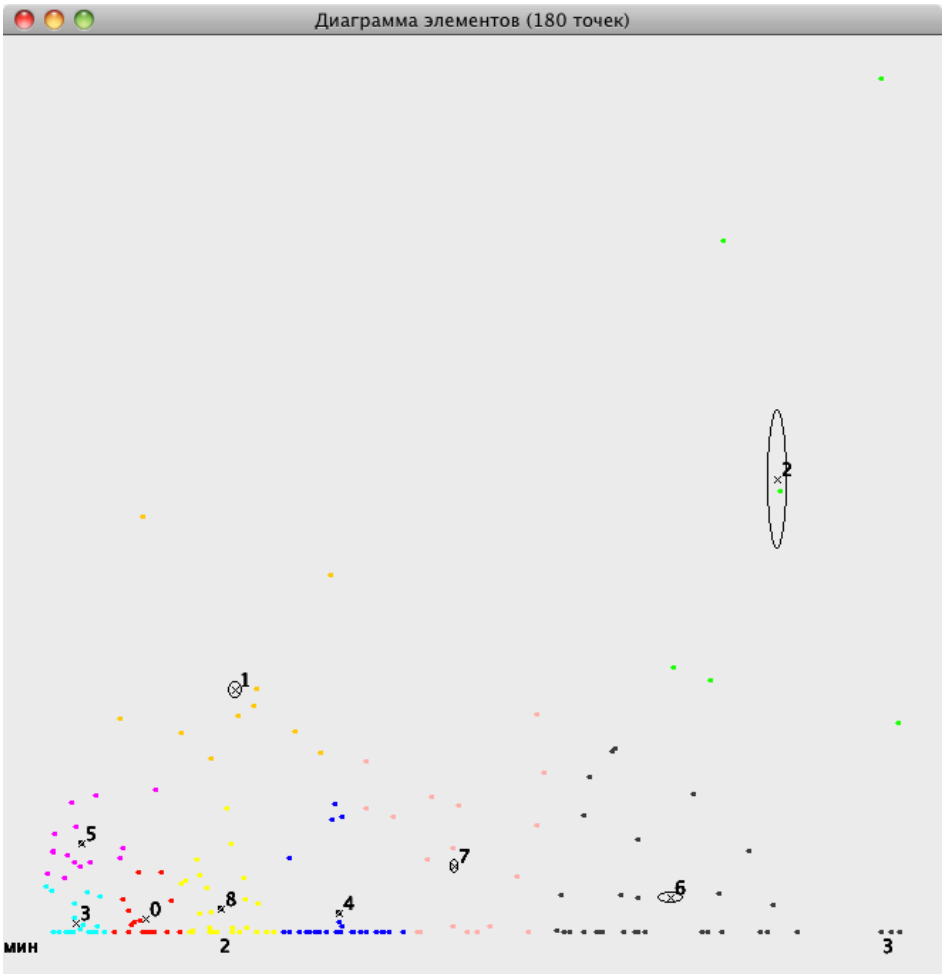

Рис. 4

Дальнейшие планы

Функционал разработанного инструмента, позволяющий проводить автоматический подсчёт статистик и кластеризацию по ним сессий, нужно будет пополнить. Ещё очень важной кажется возможность фильтрации сессий по определяемым аналитиком критериям. Хочется расширить набор способов кластеризации и применить для решения этой задачи машинное обучение.

Сейчас инструмент представляет из себя десктопное приложение и имеет из-за этого ряд существенных ограничений, связанных с производительностью и распространением полученных результатов. Одним из вариантов развития инструмента видится его перенос на отдельный веб-сервер.

Так же хочется в будущем расширить инструмент, чтобы с его помощью стало возможно анализировать большие поисковые логи.

Литература

1. Heidi Lam, Daniel Russell, Diane Tang, Tamara Munzner:
"Session Viewer: Visual Exploratory Analysis of Web Session Logs", IEEE Symposium on Visual Analytics Science and Technology, 2007. VAST 2007
2. Rui Xu, Don Wunsch “Clustering”, IEEE Press Series on Computational Intelligence