Технология выявления взаимосогласованных структур сходства пользователей и ресурсов

К. В. Воронцов, К. В. Рудаков, В. А. Лексин

Вычислительный Центр им. А. А. Дородницына РАН
voron@ccas.ru

Аннотация

Для выявления предпочтений и информационных потребностей огромного числа пользователей по отношению к огромному числу ресурсов простейшая тактика «пользователи ресурса X посещают также множество ресурсов Y» представляется не вполне адекватной. Предлагается более тонкий анализ, основанный на принципе «схожи те пользователи, которые посещают схожие множества ресурсов, и схожи те ресурсы, на которые заходят схожие пользователи». На этом принципе построена технология анализа клиентских сред (АКС), разработанная в компании Форексис [1]. В данной работе рассматривается применение АКС к обработке логов поисковой машины. Технология АКС позволяет решать задачи персонализации поиска, направленного предложения ресурсов пользователям, поиска схожих ресурсов и визуализации структуры Интернета в виде карт сходства.

1. Технология анализа клиентстких сред (АКС)

Клиентская среда — это совокупность клиентов, регулярно пользующихся некоторым фиксированным набором услуг. В последнее время всё больше компаний подробно протоколируют действия своих клиентов. Актуальной проблемой становится создание новых информационных технологий для эффективного извлечения полезных знаний из сырых данных о поведении клиентов.

Технология *анализа клиентских сред* (*АКС*) — это цепочка процедур обработки данных, ведущая от исходного протокола действий клиентов к решению широкого спектра задач маркетинга и управле-

ния взаимоотношениями с клиентами (Customer Relationship Management, CRM). К числу этих задач относятся: выявление и интерпретация типов поведения клиентов (сегментация клиентской базы), выявление целевых групп клиентов, структуризация ассортимента в соответствии с объективными предпочтениями клиентов, персонализация предложения услуг клиентам, прогнозирование возможного оттока клиентов, выявление необычного или потенциально опасного для компании поведения клиентов. Конечной целью этой деятельности является повышение качества оказываемых услуг, более эффективное привлечение и удержание клиентов.

Технология АКС основана на понятии сходства. Клиенты схожи с точки зрения компании, если они пользуются схожим набором услуг. Услуги схожи, если ими пользуются схожие множества клиентов. Данное определение приводит к паре взаимосогласованных функций сходства (*метрик*). Метрика на множестве клиентов позволяет решать задачи сегментации, поиска схожих клиентов, обнаружения необычного поведения клиентов. Метрика на множестве услуг позволяет объективно позиционировать услуги, находить сопутствующие услуги, структурировать ассортимент услуг. При решении задач персонализации услуг и направленного маркетинга приходится использовать обе метрики.

Технология АКС достаточно универсальна и может применяться в самых разных сферах бизнеса. Можно говорить о клиентских средах торговых сетей, операторов связи, организаторов биржевых торгов, эмитентов пластиковых карт, библиотек, электронных магазинов, интернет-порталов, и т. д. Например, в случае поисковой машины роль «услуг» играют ресурсы, предлагаемые в качестве результатов поиска. Клиентами являются пользователи поисковой машины. Пользование услугой — это переход клиента со страницы результатов поиска на соответствующий ресурс. Ещё один пример применения АКС — анализ результатов парламентских выборов. Здесь в качестве «услуг» выступают политические партии, «клиентами» являются регионы или избирательные участки, а «пользование услугой» соответствует тому, что на данном участке некоторый избиратель проголосовал за данную партию.

Основная технологическая цепочка АКС складывается из следующих шагов.

1. Исходными данными являются протоколы действий клиентов, в которых фиксируется: кто, когда, какой услугой, и на какую сумму воспользовался. По этим протоколам строится *частотная матрица* (называемая также матрицей кросс-табуляции), в которую записываются частоты пользования каждого клиента каждой услугой.

В зависимости от целей анализа частотная матрица может формироваться на основе объемных показателей, например, суммарной стоимости оказанных услуг.

2. На следующем этапе частотная матрица подвергается фильтрации: из нее выбрасываются клиенты и услуги с наименьшими значениями суммарной частоты пользования. Эти данные малоинформативны и только мешают выделению значимых закономерностей. Фильтрация может приводить к заметному сокращению размера частотной матрицы.

3. На основе частотной матрицы строятся две метрики — между клиентами и между услугами. Вообще говоря, существует бесконечно много различных способов построения метрик. Для выбора наиболее адекватного способа привлекаются различные дополнительные критерии.

Критерий взаимной согласованности. Повторное (итерационное) оценивание сходства клиентов через сходство услуг, и сходства услуг через сходство клиентов должно приводить к исходным метрикам на множествах клиентов и услуг.

Критерий статистической значимости. Для каждой пары клиентов (услуг) проверяется статистическая гипотеза о том, что наблюдаемое сходство является чисто случайным при заданном уровне значимости. Если это так, то считается, что надёжная информация о сходстве данной пары отсутствует.

Критерий устойчивости. Результаты анализа не должны качественно изменяться при варьировании способа вычисления метрик в некоторых разумных пределах.

Результатом данного этапа обработки являются матрицы попарных расстояний между клиентами и между услугами.

Дальнейший анализ может идти несколькими путями, в зависимости от специфики предметной области и содержательной постановки прикладных задач.

4. Кластеризация клиентов позволяет выявить структуру клиентской среды и обнаружить наиболее характерные типы клиентов. В маркетинговых исследованиях этот вид анализа называют *сегментацией клиентской базы*. Обычно сегментацию производят по совокупности агрегированных показателей, таких как средний доход от клиента, частота пользования различными категориями услуг, социально-демографические характеристики клиента, и т. д.

В АКС схожим клиентам соответствует схожий набор услуг. Это означает, что сегменты описываются в терминах наборов услуг или «типовых потребительских корзин». Такой тип сегментации полезен при маркетинговых исследованиях.

Для визуализации кластерной структуры используются либо традиционные для кластер-анализа *дендрограммы*, либо *карты сходства* — точечные графики, на которых точки соответствуют клиентам, а расстояния между точками отражают степень их сходства.

5. Проведение аналогичного анализа для услуг дает кластерную структуру ассортимента и позволяет решать задачи позиционирования услуг. Вообще, поскольку в АКС возникают сразу две взаимосогласованные метрики, практически любой анализ имеет «аналог в двойственном пространстве». Эта важная особенность АКС может приводить к полезным и порой неожиданным результатам, поскольку сама возможность постановки двойственной задачи часто ускользает от внимания при традиционных статистических методах исследования.

6. Кластеризация может быть проделана в локальном варианте — относительно заданного клиента или заданной услуги. Например, карта ближайшей окрестности некоторой услуги дает постоянному клиенту уникальную возможность быстро найти аналогичные или сопутствующие услуги, о существовании которых он, возможно, не знал, но знали похожие на него клиенты. Карта окрестности всех услуг, которыми пользовался клиент, представляет полный спектр интересов данного клиента.

7. *Направленный маркетинг* (direct marketing) — еще одна задача, решаемая с помощью АКС. Для любого конкретного клиента может быть найдено множество схожих с ним клиентов и вычислен набор наиболее востребованных среди них услуг. Если из этих услуг отбросить те, которыми данный клиент уже пользовался, получим персональное предложение, которое с большой вероятностью заинтересует данного клиента. Это и есть адресная реклама, которая, как известно, существенно более эффективна, чем массовая.

8. Набор услуг, персонально предлагаемых клиенту, может быть ранжирован в соответствии с их популярностью среди схожих клиентов, что заметно упрощает для клиента задачу выбора услуг из представленного, возможно, чрезмерно широкого списка. Применительно к клиентской среде поисковой машины это дает интересную возможность *персонализации результатов поиска*, когда ресурсы ранжируются по их популярности только среди схожих пользователей, а не среди всех пользователей Интернета.

2. Идея исследования

Цель данной работы — продемонстрировать применение общей методологии АКС к клиентской среде поисковой машины. Здесь в

роли «услуг» выступают ресурсы, предлагаемые в качестве результатов поиска. Клиентами являются пользователи поисковой машины. Пользование услугой — это переход клиента со страницы результатов поиска на соответствующий ресурс.

Анализ сходства ресурсов и клиентов позволяет предложить ряд новых сервисов и способов навигации в сети Интернет, основанных на идеях визуализации и персонализации.

3. Описание методов, алгоритмов и экспериментов

3.1. Исходные данные

Исходными данными являются протоколы переходов пользователей на ресурсы, найденные поисковой машиной. Логи поисковой машины, предоставленные компанией Яндекс для экспериментов, охватывали 7 дней, по 5–10 миллионов запросов в день. Для каждого запроса лог содержал уникальный идентификатор пользователя, список выданных документов и время обращения пользователя к выбранным им документам. Этой информации о поведении пользователей вполне достаточно для применения АКС. Тексты запросов и время обращения пользователей к документам в данном исследовании не анализировались, учитывались только сами факты выбора документов. Исследуемый лог содержал данные о 14 606 пользователях и 207 312 запросах. Из 1 972 636 документов, предлагавшихся поисковой машиной в качестве результатов поиска, 129 600 были выбраны пользователями.

3.2. Создание словарей, фильтрация пользователей и ресурсов

Основной вопрос, возникающий на начальном этапе анализа — что считать ресурсом? Неправильно было бы считать ресурсами отдельные документы, поскольку число заходов на них невелико, и оценка сходства документов, построенная по этим данным, вряд ли будет информативной. Ресурсом может быть доменное имя, однако не всегда, поскольку внутри домена может содержаться множество сайтов различной тематики, более того, отдельные сайты в свою очередь могут содержать разделы разной направленности.

Предлагается считать ресурсами достаточно часто посещаемые подкаталоги. В процессе первого чтения лога для каждого доменного имени строится дерево подкаталогов и подсчитывается число заходов в каждый каталог. Если пользователь выбрал некоторый документ, скажем, `bmw.ru/world/index.html`, то счетчик посещений увеличивается как для самого документа, так и для всех родитель-

ских каталогов, в данном случае **bmw.ru/world** и **bmw.ru**. После того, как дерево построено, выбирается порог отсечения ресурсов по числу посещений, и все узлы дерева с меньшим числом посещений отбрасываются. Оставшиеся узлы считаются ресурсами. На Рис. 1 показан пример усечения дерева ресурсов.

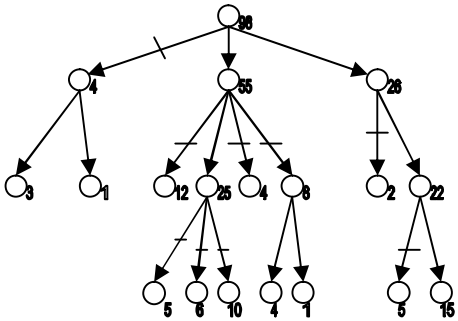

Рис. 1. Выбор ресурсов из дерева каталогов по порогу посещаемости (в данном случае 13 посещений).

Гистограмма посещаемости узлов на Рис. 2 показывает, что около 100 тысяч узлов (что составляет 77% узлов) были выбраны только один раз, около 10 тысяч — два раза, далее гистограмма очень быстро убывает. В данном исследовании был выбран порог в 30 посещений, после чего осталось 1024 ресурсов. Заметим, что увеличение количества многократно посещаемых узлов возможно путем расширения анализируемого интервала времени.

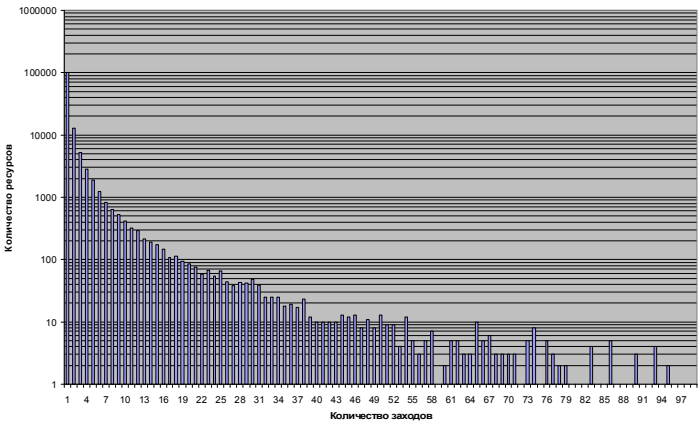

Рис. 2. Гистограмма посещаемости ресурсов

Наряду с фильтрацией ресурсов производилась фильтрация пользователей по числу посещений. При пороге в 5 посещений осталось 7292 пользователя (около половины).

Для эффективного хранения дерева ресурсов с возможностью быстрого поиска узла по имени использовалось тернарное дерево поиска [2], модифицированное таким образом, чтобы в узлах дерева хранился ещё и счетчик числа посещений. Для хранения словаря пользователей применялась хэш-таблица [3].

Этап предварительной обработки данных завершается вторым проходом логов, в ходе которого строится частотная матрица $\left\|F_{ur}\right\|_{u=1,U}^{r=1,R}$ размера $U \times R$, где U — число отобранных пользователей, R — число отобранных ресурсов, F_{ur} — количество посещений пользователем *u* ресурса *r*. Матрица *F* сильно разрежена и почти не содержит элементов, отличных от нуля и единицы. Для ее хранения используется специальная структура данных, обеспечивающая эффективный перебор и поиск ненулевых элементов в строках и столбцах [4].

3.3. Оценивание сходства ресурсов

Рассмотрим один из возможных способов оценивания сходства ресурсов, основанный на проверке статистической гипотезы о независимости посещений. Пусть *i*-й и *j*-й ресурсы посещались n_i и n_j пользователями соответственно. Пусть n_{ij} пользователей посетили оба ресурса. Если предположить, что посещения *i*-го и *j*-го ресурсов являются независимыми событиями, то случайная величина n_{ij} будет подчиняться гипергеометрическому распределению $P(x) = C_{n_i}^{x} C_{U-n_i}^{n_j - x} / C_U^{n_j}$. Эта вероятность максимальна при $x \approx n_i n_j / U$ и быстро убывает по мере увеличения *x*.

Если n_{ij} настолько велико, что $P_{ij} < \alpha$ при заданном достаточно малом уровне значимости α, где $P_{ij} = \sum_{x=n_{ij}}^{\min(n_i, n_j)} P(x)$, то можно полагать, что экспериментальные данные противоречат гипотезе о независимости ресурсов. Следовательно, имеется статистически значимая взаимосвязь в посещениях данной пары ресурсов, то есть эти ресурсы схожи. Если же $P_{ij} > \alpha$, то наблюдаемое распределение посещений (n_i, n_j, n_{ij}) вполне могло реализоваться чисто случайно.

В этом случае ресурсы не являются схожими, а значение P_{ij} можно рассматривать как неинформативный шум и полагать, что информация о сходстве данной пары ресурсов вообще отсутствует.

Введем на множестве $D_\alpha = \{(i,j) \mid P_{ij} < \alpha, i < j\}$ *функцию расстояния* между ресурсами, положив $\rho(i,j) = \mu(P_{ij})$, где μ — некоторая монотонно возрастающая функция. Чем меньше значение вероятности P_{ij}, тем более схожи *i*-й и *j*-й ресурсы.

Для выбора адекватного преобразования μ предлагается следующий подход. Фиксируется некоторое параметрическое семейство функций и производится оптимизация параметров по одному или нескольким критериям качества (критерии рассматриваются ниже). В данном исследовании использовалось двухпараметрическое семейство $\mu(P;\beta_1,\beta_2) = \beta_1 + (P/\alpha)^{\beta_2}$. Можно доказать, что в пространстве параметров (β_1,β_2) существует область значений, в которой введенная таким образом функция расстояния $\rho(i,j)$ удовлетворяет аксиомам метрики.

3.4. Многомерное шкалирование и карты сходства

Функция расстояния $\rho(i,j)$ позволяет построить *карту сходства ресурсов* — плоский точечный график, точки которого соответствуют ресурсам, а двумерные евклидовы расстояния между точками приблизительно равны исходным расстояниям $\rho(i,j)$. Данная задача решается методами *многомерного шкалирования* [5].

Пусть имеется *метрическая конфигурация* — множество *N* объектов с заданными расстояниями $\rho(i,j)$ на некотором подмножестве пар объектов $D \subseteq \{(i,j) \mid 1 \le i < j \le N\}$. Для определения координат точек (x_{i1}, x_{i2}), $i = 1,\ldots,N$, представляющих эти объекты на плоскости, решается задача минимизации *функционала стресса*:

$$S = \sum_{(i,j)\in D} w_{ij} \left(\rho(i,j) - d(i,j)\right)^2,$$

где $w_{ij} = \rho^\gamma(i,j)$ — веса объектов, $d^2(i,j) = \sum_{k=1}^{m} (x_{ik} - x_{jk})^2$ — евклидово расстояние между *i*-м и *j*-м объектами, x_{ik} — *k*-я координата точки, представляющей *i*-й объект в евклидовом пространстве размерности $m = 2$. Показатель степени γ позволяет ориентировать процесс размещения точек на более точное представление далеких

(при $\gamma > -2$) или близких (при $\gamma < -2$) расстояний. Принято считать, что наиболее адекватный результат достигается при $\gamma = -2$. В этом случае функционал стресса приобретает смысл потенциальной энергии в системе точек, соединенных упругими связями, и задача минимизации стресса приобретает физический смысл поиска устойчивого равновесия.

С решением данной оптимизационной задачи связаны две проблемы.

Во-первых, функционал стресса имеет огромное количество локальных минимумов. Ни один из известных эффективных методов многомерного шкалирования не гарантирует достижения глобального минимума стресса. Используемый в данном исследовании алгоритм не является исключением. В то же время, он ориентирован на поиск такого локального минимума, при котором сохраняются наиболее существенные структурные особенности исходной метрической конфигурации. Для построения карт сходства эта стратегия представляет даже больший интерес, чем борьба за глобальную минимизацию стресса.

Во-вторых, большинство известных алгоритмов имеют квадратичную по числу объектов сложность, что позволяет размещать до нескольких тысяч объектов за приемлемое время. Однако они практически бесполезны для сверхбольших конфигураций, насчитывающих десятки и сотни тысяч объектов. В данной работе используется алгоритм синтеза плоских представлений, имеющий субквадратичную сложность. Радикальное повышение эффективности достигается за счет того, что алгоритм просматривает не все попарные расстояния между объектами, поскольку оценки сходства $\rho(i, j)$ строятся только на подмножестве пар D_α. Кроме того, в процессе минимизации функционала стресса выявляется иерархическая кластерная структура метрической конфигурации, что позволяет ещё больше повысить эффективность.

Идею иерархического алгоритма проще всего пояснить на примере двухуровневой иерархии. В этом случае алгоритм состоит из трех этапов [6].

Первый этап — начальное размещение «скелета» из S опорных точек. Сначала размещается пара наиболее удаленных друг от друга точек. Затем к ним по очереди присоединяются другие точки. Каждая точка выбирается так, чтобы расстояние от нее до ближайшей размещенной было максимально. В процессе присоединения точки образуют все более и более мелкую сетку. Процесс продолжается, пока S точек не окажутся размещенными.

На втором этапе производится не более I «больших» итераций, в течение которых местоположения всех опорных точек уточняются поочередно. Задача второго этапа — как можно точнее выстроить скелет.

На третьем этапе размещаются все оставшиеся точки, не попавшие в число опорных. Каждая точка размещается только относительно опорных, взаимные расстояния между ними не учитываются. Поэтому время размещения конфигурации из N точек имеет порядок $O(S^2 I) + O(SN)$. Оно линейно по N, если размер скелета достаточно мал и фиксирован, и квадратично, если S имеет порядок N. Отбрасывание части информации существенно ускоряет работу алгоритма, но может несколько ухудшить качество размещения. В то же время, доказано, что если исходная метрическая конфигурация имеет *ε-кластерную структуру*, то скелет содержит по одной точке от каждого кластера, то есть правильно отражает структуру конфигурации [7],[8].

При размещении отдельной точки на всех трёх этапах используется метод Ньютона-Рафсона. Начальное приближение точки строится по трем ближайшим к ней опорным точкам. При $\gamma < 0$ именно ближайшие соседи дают основной вклад в функционал стресса, и такое начальное приближение часто оказывается близким к оптимальному. Если исходная конфигурация изначально близка к двумерной, она почти всегда будет размещена достаточно точно, даже при отключении уточняющих итераций Ньютона-Рафсона.

Описанный вариант алгоритма легко обобщается на случай иерархических структур с произвольным числом уровней. После построения скелета верхнего уровня вокруг каждой опорной точки размещается скелет второго уровня, содержащий снова не более S опорных точек. Затем строятся скелеты третьего уровня, и т. д. Скелет каждого уровня размещается только относительно точек скелета предыдущего уровня. Именно это и позволяет достичь субквадратичной эффективности алгоритма при сохранении существенных особенностей кластерной структуры метрической конфигурации.

Координаты точек, выданные алгоритмом многомерного шкалирования, используются для построения карты сходства.

Замечание 1. Карты сходства практически всегда передают исходные расстояния с некоторыми искажениями. В общем случае произвольную метрическую конфигурацию невозможно разместить на плоскости без искажений (для этого требуется, чтобы система из $N(N-1)/2$ уравнений с $2N$ неизвестными была совместна).

Замечание 2. Функционал стресса инвариантен относительно произвольных сдвигов и поворотов всей карты сходства в целом. Поэтому координатные оси не имеют содержательной интерпретации. Можно только утверждать, что точкам, близким на карте, как правило, соответствуют схожие объекты. Анализ карт сходства предполагает рассмотрение и интерпретацию плотных сгустков точек — кластеров.

3.5. Карты сходства ресурсов

На Рис. 3 приведена карта сходства всех ресурсов, построенная с помощью алгоритма многомерного шкалирования. Плотные сгустки точек на этой карте практически всегда удается четко интерпретировать, что позволяет говорить о работоспособности технологии (на рисунке отмечены только некоторые группы точек). На Рис. 4 показана внутренняя структура кластера из 37 сайтов, отмеченного на Рис. 3 как «музыка». Все они посвящены mp3-музыке, покупке компакт-дисков, плееров и т. д. Лишь 4 из них напрямую не относятся к музыке, в то же время, это 3 поисковых сайта и один Интернет-магазин, имеющие возможность поиска и покупки музыки.

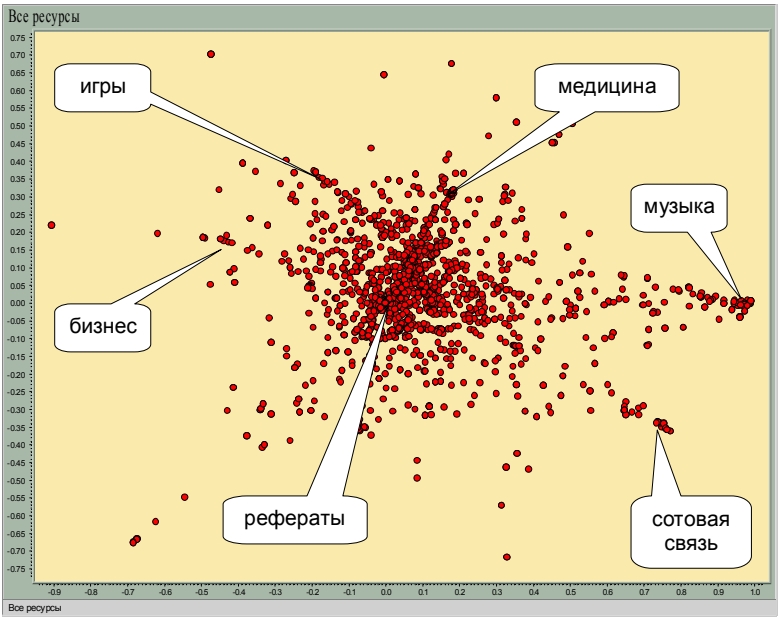

Рис. 3. Карта сходства 1024 наиболее посещаемых ресурсов.

Заметим, что на общей карте сходства в первую очередь выделяются сайты рефератов, mp3-музыки, компьютерных игр, бесплатного программного обеспечения и различных товаров: лекарств, сотовых телефонов и т. д. Таким образом, карта показывает, какие ресурсы чаще всего ищут с помощью поисковых машин, но не отражает реальных информационных потребностей массы пользователей. Более объективный взгляд на структуру российского Интернета мог бы дать анализ логов, формируемых счетчиками посещаемости.

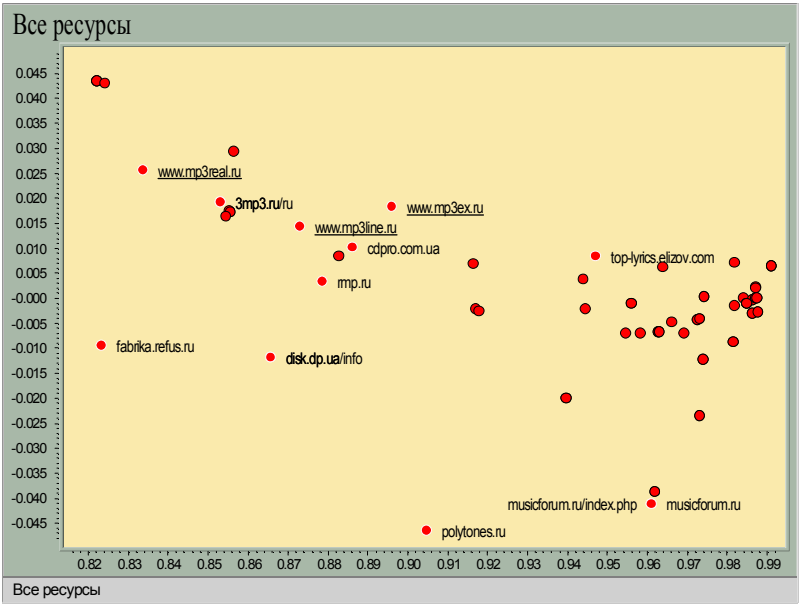

Рис. 4. Пример «плотного сгустка». Фрагмент общей карты сходства, в который попадает 37 музыкальных сайтов.

3.6. Локальные карты сходства в окрестности заданного ресурса

Окрестность ресурса r — это множество схожих с ним ресурсов $V(r) = \{r' \mid \rho(r', r) < \rho_0\}$. Порог сходства ρ_0 является параметром, который либо задается априори, либо подбирается исходя из желаемого размера окрестности $|V(r)|$. Окрестность может быть представлена графически в виде карты сходства, либо в виде списка ресурсов, ранжированных в порядке убывания сходства с ресурсом *r*.

Отметим, что *локальная* карта сходства не является простым фрагментом общей карты, как на Рис. 4. При ее построении исполь-

зуются только точки из множества $V(r)$. Поэтому локальные карты менее подвержены искажениям и гораздо быстрее строятся.

Отображение множества схожих ресурсов в виде локальной карты или ранжированного списка является ещё одним способом навигации в Интернете. Есть основания полагать, что такой сервис был бы интересен как для пользователей, так и для авторов многих сайтов. Его можно размещать в разделе полезных ссылок «Links», а также использовать для отслеживания динамики появления новых ресурсов той же тематики.

3.7. Критерии качества метрик

Построение функций сходства (метрик) имеет ряд преимуществ перед стандартной методикой «пользователи ресурса *X* посещают также множество ресурсов *Y*». Метрики позволяют применить богатый арсенал методов кластеризации, распознавания образов и многомерного шкалирования. Однако результаты их применения существенно зависят от качества используемой метрики. Как правило, понятие сходства допускает множество различных формализаций, поэтому возникает не только необходимость, но и возможность оптимизации метрики.

Рассмотрим некоторые критерии качества, применяемые в технологии АКС для оптимизации метрик.

Критерий интерпретируемости функции сходства. Формируется выборка ресурсов с известным заранее разбиением на классы — темы. Ресурсы, относящиеся к нескольким темам, исключаются из выборки. Критерий определяется как отношение среднего внутриклассового расстояния к среднему межклассовому расстоянию:

$$R(\rho) = \frac{|D_\alpha^{\text{out}}|}{|D_\alpha^{\text{in}}|} \frac{\sum\limits_{(i,j)\in D_\alpha^{\text{in}}} \rho(i,j)}{\sum\limits_{(i,j)\in D_\alpha^{\text{out}}} \rho(i,j)},$$

где $D_\alpha^{\text{in}} = \{(i,j) \in D_\alpha \mid K_i = K_j\}$, $D_\alpha^{\text{out}} = \{(i,j) \in D_\alpha \mid K_i \neq K_j\}$, K_i — номер класса, к которому относится *i*-й ресурс. Разумеется, невозможно каждую выборку заранее классифицировать вручную. Оптимизация параметров метрики и уровня значимости α производится один раз по обучающей выборке, сформированной описанным выше способом.

Когда результаты представляются в виде карт сходства, более адекватным является *критерий интерпретируемости карты сходства*. Он строится так же, как в предыдущем случае, с тем отличи-

ем, что оценивается не исходная функция сходства $\rho(i,j)$, а результат шкалирования — плоская евклидова метрика $d(i,j)$. Таким образом, данный критерий одновременно оценивает как интерпретируемость метрики, так и ее *планарность* — возможность «спроецировать» метрическую конфигурацию на плоскость с минимальными искажениями.

Критерий нормированного стресса оценивает относительную величину искажений, возникающих в результате проецирования:

$$S(\rho) = \frac{\sum\limits_{(i,j)\in D_\alpha} w_{ij} \left|\rho(i,j) - d(i,j)\right|}{\sum\limits_{(i,j)\in D_\alpha} w_{ij}\rho(i,j)}.$$

Критерий метричности оценивает, является ли функция сходства метрикой (в этом случае функционал принимает наименьшее значение, равное нулю), и если не является, то насколько сильно нарушаются неравенства треугольника:

$$T(\rho) = \frac{1}{|T_\alpha|} \sum_{(i,j,k)\in T_\alpha} \left(\frac{\rho(i,j)}{\rho(i,k)+\rho(k,j)} - 1 \right)_+ ;$$

$$T_\alpha = \left\{ (i,j,k) \,\middle|\, (i,j)\in \tilde{D}_\alpha, (i,k)\in \tilde{D}_\alpha, (k,j)\in \tilde{D}_\alpha \right\};$$

$$\tilde{D}_\alpha = D_\alpha \cup \left\{ (i,j) \,\middle|\, (j,i)\in D_\alpha \right\}.$$

Несколько критериев могут использоваться одновременно. Например, если некоторые параметры сильнее влияют на метричность, а другие — на интерпретируемость, то имеет смысл оптимизировать эти параметры по отдельности.

3.8. Оценка полезности ресурсов для пользователя

Метрики на множестве пользователей можно строить точно так же, как это было сделано выше для ресурсов. Методы кластеризации, шкалирования и визуализации в пространстве пользователей применяются в маркетинговых исследованиях при сегментации клиентской базы. Здесь мы рассмотрим ещё одно применение — построение оценок полезности ресурсов для заданного пользователя и генерацию персонального предложения (*директ-маркетинг*).

Окрестность пользователя u — это множество схожих с ним пользователей $V(u) = \{u' \mid \rho(u',u) < \rho_0\}$. Порог сходства ρ_0 может задаваться априори, либо подбираться исходя из желаемого размера окрестности $|V(u)|$.

Оценка полезности H_{ur} *ресурса* r *для пользователя* u — это средняя взвешенная посещаемость ресурса r схожими с ним пользователями из множества $V(u)$:

$$H_{ur} = \frac{\sum_{v \in V(u)} K(\rho(u,v)) F_{vr}}{\sum_{v \in V(u)} K(\rho(u,v))},$$

где F_{vr} — количество посещений пользователем v ресурса r, $K(\rho)$ — неотрицательная монотонно убывающая функция, удовлетворяющая условию $K(0) = 1$. Веса пользователей $w(v) = K(\rho(u,v))$ убывают по мере увеличения расстояния до u.

3.9. Персональное предложение (директ-маркетинг)

Для формирования *персонального предложения* вычисляются оценки полезности всех ресурсов для заданного пользователя u и составляется ранжированный список наиболее полезных ресурсов. Результат можно также представить графически в виде локальной карты сходства.

Альтернативный подход к формированию персонального предложения заключается в использовании метрики на множестве ресурсов. Строится множество ресурсов, схожих с ресурсами $r_1, \dots r_k$, уже посещенными пользователем u:

$$V(r_1, \dots r_k) = \left\{ r' \middle| \min_{1 \le i \le k} \rho(r', r_i) < \rho_0 \right\}.$$

Это множество представляется в виде карты сходства и/или списка, ранжированного по сходству с ближайшим посещённым ресурсом.

На Рис. 5 показан пример локальной карты сходства с персональным предложением ресурсов для одного из пользователей. На Рис. 6 показан фрагмент этой карты, который еще раз свидетельствует о хорошей интерпретируемости карт сходства — плотная группа точек на карте практически полностью состоит из сайтов одинаковой тематики (в данном случае — рефератов).

3.10. Персонализация результатов поиска

Еще одно применение оценок полезности ресурсов — *персонализация результатов поиска*. Пусть имеется пользователь, обратившийся с запросом к поисковой машине, и список найденных документов. Идея метода заключается в том, чтобы передвигать ближе к началу списка те документы, которые имеют более высокую оценку полезности для данного пользователя.

Чем больше данных накоплено о поведении пользователя, тем более объективным будет ранжирование списка. Поиск автоматически будет настраиваться на выдачу ресурсов, близких к тем, которые обычно вызывают интерес у данного пользователя.

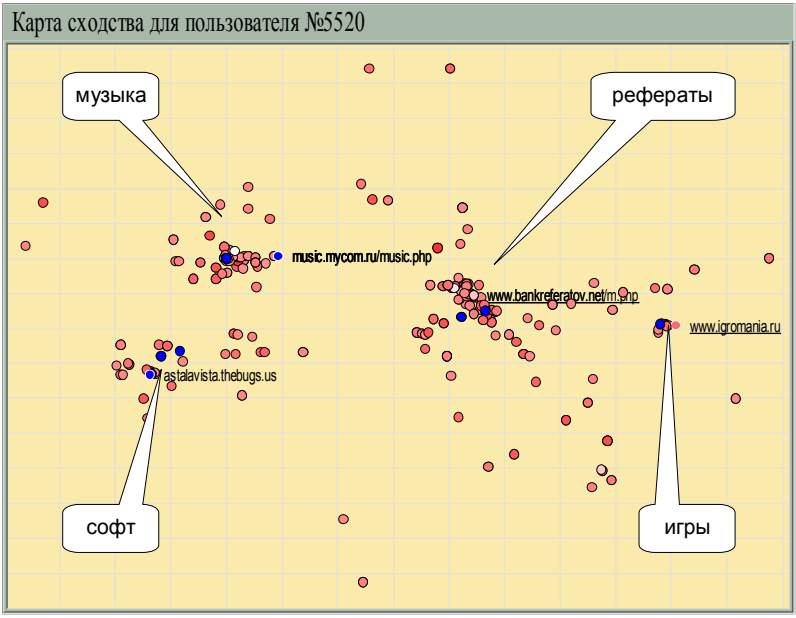

Рис. 5. Карта сходства ресурсов, предлагаемых конкретному пользователю, который посещал сайты по четырем темам: музыка, рефераты, компьютерные игры, бесплатный софт. Посещенные сайты отмечены синими точками. Соответствующим образом выстроилось и персональное предложение.

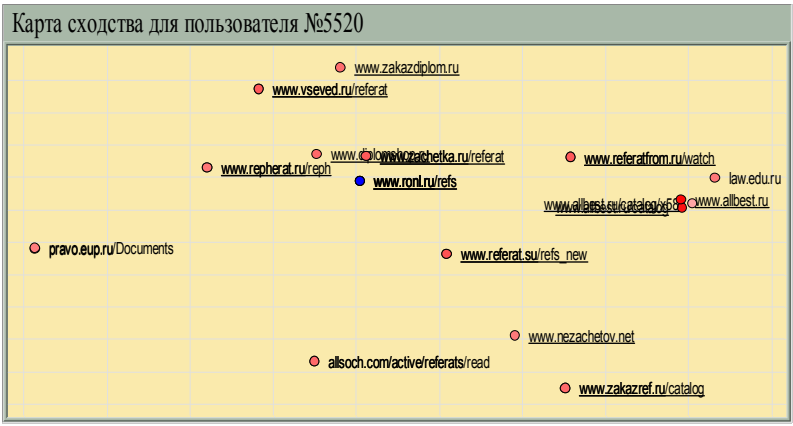

Рис. 6. Увеличенный фрагмент карты ресурсов, предложенных тому же пользователю — сайты с рефератами.

3.11. Поиск схожих пользователей

Еще одно любопытное потенциальное применение метрики на множестве пользователей — поиск схожих пользователей. Сервис мог бы быть реализован следующим образом. Зарегистрированные пользователи сообщают некоторую информацию о себе и соглашаются с тем, что за их поведением в Интернете «будут следить». Каждый зарегистрированный пользователь имеет свою страницу, на которой он видит список ближайших к нему пользователей, предоставленную ими информацию о себе и автоматически сформированный список интересующих их тем и ресурсов. Дальнейшие контакты пользователей со своими «единомышленниками» возлагаются на самих пользователей.

4. Выводы и обсуждение результатов

Технология анализа клиентских сред (АКС) основана на понятии сходства. Ресурсы схожи, если на них заходят схожие множества пользователей. Пользователи схожи, если они посещают схожие множества ресурсов.

Для определения сходства применяется техника проверки статистических гипотез. Однако это ещё не позволяет однозначно ввести метрику в пространстве ресурсов (или пользователей). Для ведения метрики используются параметрические семейства монотонных отображений, параметры которых настраиваются по критериям интерпретируемости, метричности, планарности и др.

Наличие метрики позволяет применять богатый арсенал методов кластерного анализа, распознавания образов, непараметрической статистики. В частности, методы многомерного шкалирования позволяют строить наглядное представление кластерной структуры ресурсов в виде карт сходства.

Карты сходства, построенные по большому количеству ресурсов, можно использовать как средство графической визуализации тематической структуры Интернета, а также как оригинальное средство навигации. Несмотря на неизбежные искажения расстояний, возникающие при построении карт сходства, в целом они неплохо отражают кластерную структуру множества ресурсов. Плотные группы точек на карте практически всегда удается интерпретировать как сайты схожей тематики. В то же время, карта сходства, построенная по логам поисковой машины, отражает скорее то, какие ресурсы чаще ищут с помощью поисковых машин, чем реальные информационные потребности массы пользователей. Более объективный взгляд на структуру российского Интернета мог бы дать анализ логов, формируемых счетчиками посещаемости.

Анализ окрестности конкретного ресурса позволяет автоматически генерировать локальные карты сходства или списки близких ресурсов, которые могут представлять значительный интерес как для пользователей, так и для создателей данного ресурса.

Совместное применение двух метрик — на множестве ресурсов и на множестве пользователей — позволяет генерировать направленное предложение ресурсов пользователям (директ-маркетинг).

Наличие метрики на множестве пользователей позволяет оценивать полезность любого ресурса для любого пользователя. Одно из возможных применений этих оценок — персонализация результатов поиска. Идея заключается в том, чтобы ранжировать ресурсы по их популярности среди схожих пользователей, а не среди всех пользователей Интернета.

Еще одно любопытное потенциальное применение метрики на множестве пользователей — автоматизация поиска «единомышленников» — пользователей, схожих по своему поведению в Интернете.

Литература

[1] Технология анализа клиентских сред. Форексис. 2005. http://www.forecsys.ru/cea.php

[2] Bentley J., Sedgewick B. Ternary Seach Trees. Dr. Dobb's Journal. April 1998. http://www.ddj.com/documents/s=921/ddj9804a

[3] Loudon K. Mastering Algorithms with C. First Edition. 1999. Chapter 8: Hash Tables.
http://www.oreilly.com/catalog/masteralgoc/chapter/ch08.pdf
[4] Tran Van Canh. Represent sparse matrices by some appropriate form of linked lists. 2002.
http://www.codeproject.com/cpp/sparse_matrices.asp
[5] Дэйвисон М. Многомерное шкалирование. Методы наглядного представления данных. Москва. Финансы и статистика, 1988.
[6] Воронцов К. В., Вальков А. С. О быстрых алгоритмах синтеза плоских представлений метрических конфигураций. Искусственный Интеллект, Донецк, 2004. №2 с.43–48.
[7] Вальков А. С. О быстрых алгоритмах синтеза плоских представлений конечных метрических конфигураций // Ж. вычисл. матем. и матем. физ. 2005. Т.45. №2, с.357–368.
[8] Вальков А. С. О быстром алгоритме восстановления иерархической ε-кластерной структуры при $\varepsilon < 1$ // Ж. вычисл. матем. и матем. физ. 2005. Т.45. №1, с.170–179.

Mining similarity structures of users and web sites

K. V. Vorontsov, K. V. Rudakov, V. A. Leksin

There are some well-known limitations for personalization techniques like collaborative filtering based on users similarity. In this paper a more general framework of Customer Environment Analysis, CEA is considered that was first introduced by FORECSYS (www.forecsys.ru). The CEA technology is based on a pair of dual similarities—between clients (users) and between objects (goods, resources, cites, etc.): “Clients are similar if they use similar sets of objects. Objects are similar if they are used by similar sets of clients.” This technology is applied to analyze a log-file of Yandex search engine. We show how CEA helps to solve different analytical tasks: personalization, direct marketing, segmentation and visualization of Internet segments by means of multidimensional scaling and Similarity Maps.