УДК 519.237
К. В. Воронцов, К. В. Рудаков, В. А. Лексин, А. Н. Ефимов
Вычислительный центр РАН,
г. Москва, Россия
voron@ccas.ru

ВЫЯВЛЕНИЕ И ВИЗУАЛИЗАЦИЯ МЕТРИЧЕСКИХ СТРУКТУР НА МНОЖЕСТВАХ ПОЛЬЗОВАТЕЛЕЙ И РЕСУРСОВ ИНТЕРНЕТ

АННОТАЦИЯ

В статье рассматривается применение технологии Анализа Клиентских Сред (АКС) для выявления, визуализации и практического использования информации о сходстве сайтов и пользователей сети Интернет на основе данных о посещениях.

О ТЕХНОЛОГИИ АНАЛИЗА КЛИЕНТСКИХ СРЕД

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. Разработанная авторами технология анализа клиентских сред, АКС [1] — это цепочка процедур обработки данных, ведущая от исходного протокола действий клиентов к решению широкого спектра задач принятия решений, маркетинга и управления взаимоотношениями с клиентами (CRM).

Технология АКС основана на понятии сходства. Клиенты схожи, если они пользуются схожими сервисами. Сервисы схожи, если ими пользуются схожие клиенты. Данное определение рекурсивно и приводит к паре взаимосогласованных метрик. Метрика на множестве клиентов позволяет решать задачи сегментации клиентской базы, поиска схожих клиентов, обнаружения необычного поведения клиентов. Метрика на множестве сервисов позволяет структурировать ассортимент, объективно позиционировать сервисы, находить сопутствующие и взаимозаменяемые сервисы. При решении задач персонализации и направленного маркетинга используются обе метрики.

Технология АКС универсальна и может применяться в торговых сетях, телекоммуникационных компаниях, Интернет-магазинах, и многих других сферах.

О ТЕХНОЛОГИИ АНАЛИЗА КЛИЕНТСКИХ СРЕД

В данном исследовании АКС применяется для анализа поведения пользователей в сети Интернет (Web Usage Mining). Рассматривается клиентская среда поисковой машины. Роль «сервисов» играют страницы или документы, предлагаемые в качестве результатов поиска. «Клиентами» являются пользователи поисковой машины. «Пользование сервисом» — это переход клиента со страницы результатов поиска на соответствующий ресурс. Исходными данными являются протоколы переходов пользователей. Протоколы, предоставленные компанией Яндекс, охватывали 7 дней, по 5–10 миллионов запросов в день. Для каждого запроса в протоколе фиксировался уникальный идентификатор пользователя, список выданных документов и время обращения пользователя к выбранным документам.

Основной вопрос, возникающий на начальном этапе анализа — что считать ресурсом? Неправильно было бы считать ресурсами отдельные документы, поскольку число заходов на них невелико, и оценки сходства для них вряд ли будут информативными. Ресурсом может быть доменное имя, однако не всегда, поскольку внутри домена часто находятся сайты различной тематики. Предлагается каждый сайт представлять в виде дерева каталогов и рассматривать в качестве ресурсов только достаточно посещаемые узлы дерева. Для пользователей также выставляется порог минимального числа посещений.

После фильтрации ресурсов и пользователей строится частотная матрица, содержащая количество посещений каждого ресурса каждым пользователем. Затем по этой матрице вычисляются попарные расстояния между ресурсами и между пользователями. Оценка сходства ресурсов основывается 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 значение n_{ij} настолько велико, что вероятность его реализации $P_{ij} < \alpha$ при заданном достаточно малом уровне значимости α, то можно полагать, что экспериментальные данные противоречат гипотезе независимости. Посещения данной пары ресурсов статистически взаимосвязаны, следовательно, эти ресурсы схожи. Если же $P_{ij} > \alpha$, то наблюдаемое распределение посещений (n_i, n_j, n_{ij}) вполне могло реализоваться чисто случайно, и ре-

сурсы нельзя считать схожими. Тогда значение P_{ij} является неинформативным шумом, и можно полагать, что информация о сходстве данной пары ресурсов вообще отсутствует. Расстояние между ресурсами определяется как $\rho(i, j) = \mu(P_{ij})$, где μ — монотонно возрастающая функция, подбираемая по критериям качества кластеризации и минимума суммарных искажений на картах сходства (см. ниже). Аналогично строится матрица расстояний между пользователями, причём обе матрицы оказываются сильно разреженными. Разреженность является полезным свойством, так как позволяет строить эффективные алгоритмы, основанные на анализе локальных окрестностей в метрических пространствах ресурсов и пользователей. Дальнейший анализ может идти несколькими путями, в зависимости от содержательной постановки прикладных задач.

КАРТЫ СХОДСТВА

Метрика на множестве ресурсов позволяет применить иерархические методы многомерного шкалирования для визуализации множества ресурсов в виде *карт сходства* [2]. Точками на карте отображаются ресурсы; при этом близкие точки, как правило, соответствуют схожим ресурсам. На Рис. 1 приведена карта сходства 1024 ресурсов.

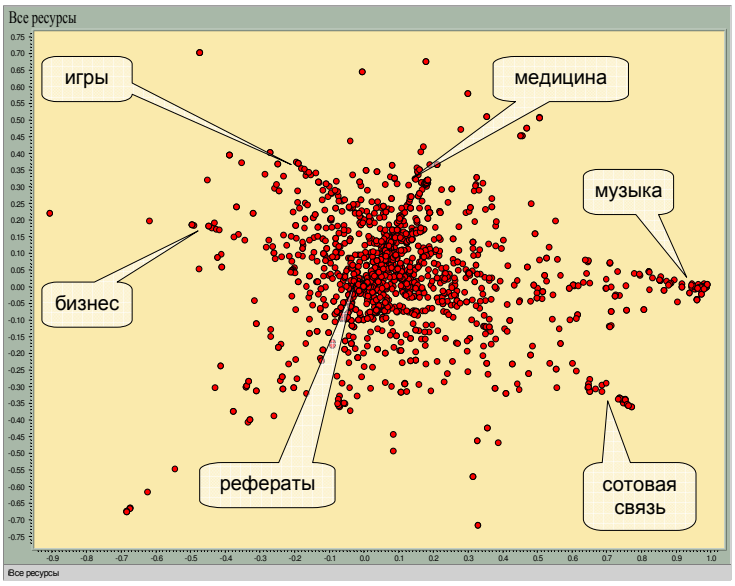

Рис. 1. Карта сходства всех ресурсов.

Несмотря на неизбежные искажения расстояний, возникающие при многомерном шкалировании, карты сходства адекватно отражают тематическую кластерную структуру ресурсов. Плотные группы точек на карте практически всегда удаётся интерпретировать как сайты схожей тематики (на Рис. 1 отмечены лишь некоторые группы сайтов), несмотря на то, что при вычислении оценок сходства содержимое ресурсов не учитывается, а используются только данные о посещениях. Карты сходства можно рассматривать как своеобразное средство навигации. Каждая точка представляет собой ссылку на соответствующий сайт. Функция интерактивного масштабирования позволяет детально исследовать отдельные группы точек.

Локальные карты сходства позволяют визуализировать подмножества ресурсов, близких к заданному ресурсу или группе ресурсов. Одно из возможных практических применений — автоматическая генерация карты схожих сайтов для традиционного раздела полезных ссылок «Links». Другое применение — визуализация направленного предложения ресурсов для заданного пользователя, см. Рис 2.

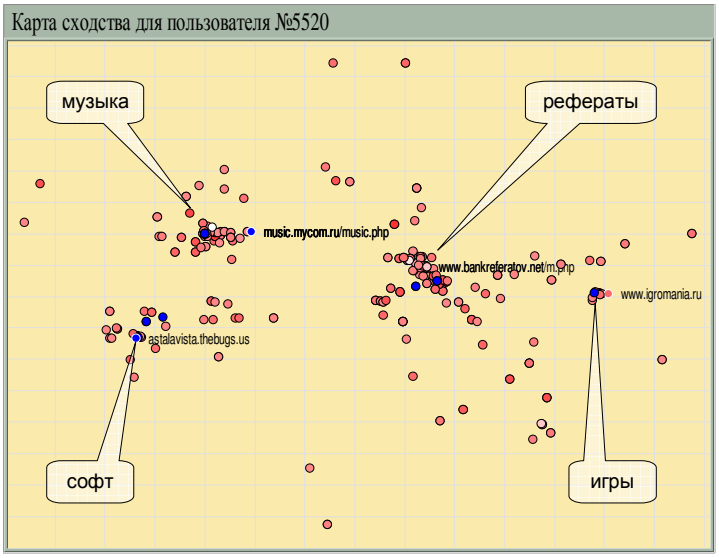

Рис. 2. Карта сходства ресурсов, предлагаемых конкретному пользователю.
Различаются сайты, посещённые пользователем, и сайты, предлагаемые пользователю.

ЗАКЛЮЧЕНИЕ

Применение технологии АКС для анализа данных о посещаемости открывает новые возможности, как перед пользователями, так и перед авторами Интернет-сайтов.

Графическая навигация: глобальные и локальные интерактивные карты сходства сайтов являются новым оригинальным средством навигации в сети Интернет.

Сегментация пользователей: выявление информационных предпочтений пользователей позволяет лучше понимать целевую аудиторию сайтов, занимающихся электронной коммерцией.

Автоматическое создание каталогов: оценки сходства ресурсов позволяют проводить иерархическую кластеризацию и каталогизацию ресурсов.

Направленное предложение (direct marketing): предложение клиенту только тех ресурсов или услуг, которые наиболее востребованы среди схожих клиентов.

Персонализация результатов поиска: ранжирование результатов поисковых запросов по популярности только среди схожих пользователей.

Создание адаптивных сайтов: автоматическая настройка внешнего вида сайта и рекламного контента с учетом персональных предпочтений пользователя.

Поиск единомышленников: поиск схожих пользователей и построение карт пользовательских сообществ на веблогах и форумах.

Работа выполнена при поддержке РФФИ (проект №05-07-90410) и ООО «Яндекс», www.yandex.ru.

ЛИТЕРАТУРА

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

2. Воронцов К. В., Вальков А. С. О быстрых алгоритмах синтеза плоских представлений метрических конфигураций. Искусственный Интеллект, Донецк, 2004. №2 с.43–48.

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

Виявлення і візуалізація метричних структур на безлічах пользователей і ресурсів Інтернет

У статті розглядається застосування технології Аналізу Клієнтських Середовищ (АКС) для виявлення, візуалізації і практичного використання інформації про подібність сайтов і користувачів мережі Інтернет на основі даних про відвідування.

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

Web Usage Mining based on web users and web sites similarity measures

A Client Environment Analysis (CEA) technology is considered with application to Web Usage Mining. Our approach is based on similarity estimation between web users and between web sites from raw data on users' behavior.