УДК 519.25
В. А. Лексин
Московский физико-технический институт (государственный университет)

ПЕРСОНАЛИЗАЦИЯ КОНТЕНТА НА ОСНОВЕ ОЦЕНОК СХОДСТВА ПОЛЬЗОВАТЕЛЕЙ И РЕСУРСОВ СЕТИ ИНТЕРНЕТ

Web usage mining (WUM) — это активно развивающаяся область исследований, направленных на разработку эффективных технологий для извлечения полезной и нетривиальной информации из протоколов посещений ресурсов пользователями сети Интернет. Типовые задачи WUM — выявление информационных предпочтений пользователей, сегментация пользователей и/или ресурсов для маркетинговых исследований и электронной коммерции, персонализация контента и направленная реклама.

В данной работе рассматривается подход к анализу протоколов, основанный на вычислении согласованных оценок сходства как между пользователями, так и между ресурсами, по принципу «пользователи схожи, если они посещают схожие множества ресурсов; ресурсы схожи, если их посещают схожие пользователи».

По исходным протоколам строится матрица кросс-табуляции $F = \|f_{ur}\|_{U\times R}$, в которой f_{ur} характеризует объём посещения r-го ресурса u-ым пользователем. На основе матрицы F строятся две матрицы попарных расстояний — между пользователями и между ресурсами. Оценка сходства пары ресурсов основывается на проверке статистической гипотезы о независимости посещений. Чем меньше вероятность чисто случайной реализации наблюдаемого числа пользователей, посетивших оба ресурса, тем меньше расстояние между ресурсами. Если же вероятность превышает заданный уровень значимости, то предполагается, что информация о сходстве отсутствует. Реализация этого принципа приводит к построению сильно разреженных матриц расстояний, что позволяет применять высокоэффективные алгоритмы.

Персонализация контента — это представление каждому пользователю наиболее интересной для него информации в наиболее удобном для него виде. Технология персонализации предложения ресурсов заданному целевому пользователю на основе мер сходства предполагает выполнение следующих шагов:

1) поиск пользователей, схожих с целевым пользователем;

2) выделение списка ресурсов, посещавшихся схожими пользователями;

3) пополнение списка ресурсами, схожими с выделенными;

4) сортировка рекомендуемого списка.

Для построения качественного персонального предложения необходимо убедиться в том, что используемые меры сходства адекватны. Поскольку формального критерия качества не существует, применялся визуальный анализ: методом многомерного шкалирования строились карты сходства ресурсов [1]. Хотя для вычисления сходства использовались только данные о посещениях, близкими на картах оказывались, как

правило, ресурсы схожей тематики. Это как раз и свидетельствует об адекватности построенных мер сходства. На Рис. 1 показана карта сходства ресурсов, построенная по данным поисковой машины, предоставленным ООО «Яндекс» (www.yandex.ru). Тематическая раскраска карты основана на априорной частичной классификации ресурсов и применении диаграммы Вороного.

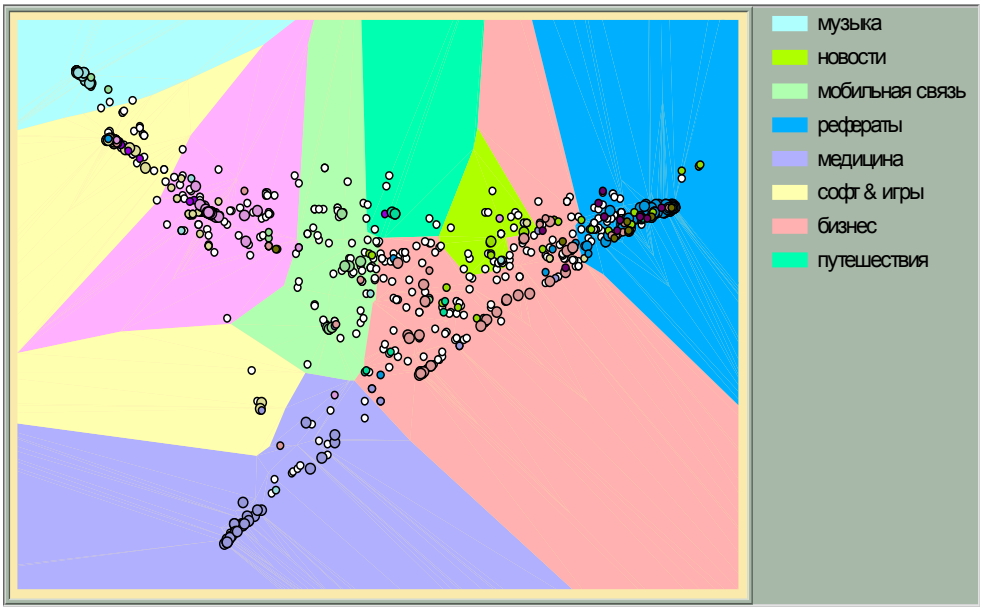

Рис. 1. Карта сходства ресурсов

Таким образом, предлагаемая технология персонализации позволяет не только формировать списки ресурсов, но и представлять их в наглядной визуальной форме. В частности, строить карты сходства ресурсов, рекомендуемых данному пользователю, а также ресурсов, схожих с данным ресурсом, и использовать эти карты, фактически, как средство навигации в сети Интернет.

Работа выполнена при поддержке РФФИ (проект №05-07-90410).

Литература

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