УДК 519.25

Оценивание сходства пользователей и ресурсов путем выявления скрытых тематических профилей

Рассматривается задача Анализа Клиентских Сред, состоящая в следующем. Заданы множество пользователей *U*, множество ресурсов *R*, и имеются данные о посещениях в виде множества пар $D = (u_i, r_i)_{i=1}^{l} \subset U \times R$. Требуется построить функции сходства на множествах пользователей $\rho_U(u,u')$ и ресурсов $\rho_R(r,r')$.

Особенность предложенного в работе подхода заключается в оценивании скрытых тематических профилей пользователей и ресурсов [4]. Предположим, что существует множество информационных потребностей *T*, называемых далее *темами*. Профилем пользователя $u \in U$ назовем вектор вероятностей $p_{tu} = p(t \mid u)$ того, что данный пользователь интересуется темой *t*, причём $\sum_{t \in T} p_{tu} = 1$. Профилем ресурса $r \in R$ назовем вектор вероятностей $q_{tr} = q(t \mid r)$ того, что данный ресурс удовлетворяет теме *t*, где $\sum_{t \in T} q_{tr} = 1$. Задача заключается в том, чтобы по протоколу *D* восстановить неизвестные тематические профили пользователей $\{p_{tu}, t \in T\}, u \in U$ и ресурсов $\{q_{tr}, t \in T\}, r \in R$. Распишем вероятность выбора пользователем *u* ресурса *r* двумя различными способами:

$$p(u,r) = \sum_{t \in T} p(u) p_{tu} q(r \mid t,u) = \sum_{t \in T} \left(p(u) p_{tu} q_{tr} q(r) / \sum_{r' \in R} q_{tr'} q_{r'} \right), \quad (1)$$

$$p(u,r) = \sum_{t \in T} q(r) q_{tr} p(u \mid t,r) = \sum_{t \in T} \left(q(r) q_{tr} p_{tu} p(u) / \sum_{u' \in U} p_{tu'} p_{u'} \right), \quad (2)$$

где $p(u)$ и $q(r)$ — априорные вероятности появления клиента *u* и ресурса *r* в записи протокола, $q(r \mid t,u)$ и $p(u \mid t,r)$ — апостериорные вероятности, которые в (1) и (2) выражены по формуле Байеса через профили ресурсов и пользователей.

Для нахождения профилей применим принцип максимума правдоподобия:

$$\sum_{i=1}^{l} \ln p(u_i, r_i) \to \max,$$

где максимум берется по всем профилям $\{p_{tu}\}$, $\{q_{tr}\}$ при ограничениях типа равенства $\sum_{t \in T} p_{tu} = 1$ для всех $u \in U$ и $\sum_{t \in T} q_{tr} = 1$ для всех $r \in R$. Для решения оптимизационной задачи предлагается итерационный процесс, в котором повторяются два шага:

1) оптимизация профилей $\{p_{tu}\}$ при фиксированных профилях $\{q_{tr}\}$;

2) оптимизация профилей $\{q_{tr}\}$ при фиксированных профилях $\{p_{tu}\}$.

Оптимизация на каждом шаге выполняется с помощью EM-алгоритма. Скрытыми переменными в EM-алгоритме являются апостериорные вероятности того, что

пользователь u, зайдя на ресурс r, удовлетворяет свой интерес t. Главная особенность алгоритма заключается в том, что использование симметричных разложений (1) и (2) обеспечивает взаимную согласованность профилей пользователей и ресурсов.

Для оптимизации параметров алгоритма и исследования сходимости был произведен эксперимент на модельных данных при $|R| = 300$, $|U| = 600$, $|T| = 10$, $l = 10\,000$. Истинные профили задавались случайным образом, по 2 темы в каждом профиле. Для генерации выборки посещений использовалась та же вероятностная модель (1). Точность восстановления профилей оценивалась по среднеквадратичному отклонению от истинных профилей. Выяснилось, что 6 итераций на внешнем цикле и 2 EM-итераций на внутреннем цикле вполне достаточно для восстановления, причём дальнейшее увеличение числа внутренних итераций даже немного ухудшает точность.

Искомые расстояния $\rho_U(u,u')$ и $\rho_R(r,r')$ определяются как евклидовы расстояния между тематическими профилями. Это способ сравнивался со стандартными методами оценивания расстояний через корреляцию [2] и через вероятность случайного независимого совместного выбора [1]. Для сравнения трёх метрик был проведён эксперимент на протоколах поисковой машины Яндекс. Исходными данными являлись протоколы переходов пользователей со страниц результатов поиска. Качество метрик оценивалось как доля ресурсов, ошибочно классифицированных методом k ближайших соседей (после выбора оптимального k). В качестве обучающей выборки использовалась классификация 396 ресурсов на 8 тематических классов. Оказалось, что доля ошибок классификации для предложенного метода составляет 11%, для метрики по вероятности случайного совместного выбора – 25%, для метрики через корреляцию – 38%.

Работа выполнена при поддержке РФФИ, проекты №05-07-90410, 07-01-12076-офи.

Литература

1. *Воронцов К. В., Рудаков К. В., Лексин В. А., Ефимов А. Н.* Выявление и визуализация метрических структур на множествах пользователей и ресурсов Интернет // Искусственный Интеллект, Донецк, 2006. — №2. — С. 285–288.

2. *Resnick P., Iacovou N., Suchak M., Bergstorm P., Riedl J.* GroupLens: an open architecture for collaborative filtering of Netnews // proc. of ACM conf. on Computer supported cooperative work, 1994. — Pp. 175–186.

3. *Schein A., Popescul A., Ungar L., Pennock D.*. Generative models for cold-start recommendations // SIGIR'01 Workshop on Recommender Systems, 2001.

4. *Воронцов К. В., Лексин В. А.* Анализ клиентских сред: выявление скрытых профилей и оценивание сходства клиентов и ресурсов // Математические методы распознавания образов: 13-ая Всеросс. конф.: Докл. М.: МАКС Пресс, 2007. С. 488–491.