

## Анализ клиентских сред: выявление скрытых профилей и оценивание сходства клиентов и ресурсов

*Воронцов К. В., Лексин В. А.*

Москва, ЗАО «Форексис», МФТИ  
vleksin@mail.ru, voron@ccas.ru

Клиентская среда — это совокупность клиентов, регулярно пользующихся фиксированным набором ресурсов (услуг, товаров, сервисов). Клиентскими средами обладают торговые сети, операторы связи, интернет-магазины, поисковые машины, электронные библиотеки, эмитенты пластиковых карт, и т. д.

Анализ клиентских сред (АКС) — это технология обработки исходных данных о действиях клиентов, направленная на решение таких задач, как персонализация предложений клиентам, поиск схожих ресурсов, каталогизация ресурсов, сегментация клиентской базы, формирование клиентских сообществ, выявление нетипичного поведения клиентов, и т. д. Технология АКС основана на вычислении оценок сходства между ресурсами и между клиентами согласно *принципу согласованности*: «ресурсы схожи, если ими пользуются схожие клиенты; в то же время, клиенты схожи, если они пользуются схожими ресурсами» [1].

В данной работе предлагается новый алгоритм АКС, основанный на выявлении интересов (скрытых профилей) клиентов и ресурсов. Алгоритм напоминает генеративные (generative model based) методы коллаборативной фильтрации (collaborative filtering) [2], но отличается от них применением принципа согласованности.

### Восстановление профилей клиентов и ресурсов

Пусть заданы:  $U$  — множество клиентов,  $R$  — множество ресурсов,  $D = (u_i, r_i)_{i=1}^{\ell} \subset U \times R$  — выборка (протокол) действий клиентов. Пара  $(u_i, r_i)$  означает, что клиент  $u_i$  воспользовался ресурсом  $r_i$  в момент времени  $i$ . Задано множество возможных интересов или тем  $T$ . Допустим, что каждый клиент  $u \in U$  интересуется темой  $t \in T$  с вероятностью  $p_{tu} = p(t|u)$ , и каждый ресурс  $r \in R$  способен удовлетворить интерес  $t$  с вероятностью  $q_{tr} = q(t|r)$ . Здесь и далее все вероятности, относящиеся к ресурсам, обозначаются буквой  $q$ . Задача состоит в том, чтобы по протоколу  $D$  восстановить *профили клиентов*  $\{p_{tu} : t \in T\}$ ,  $u \in U$  и *профили ресурсов*  $\{q_{tr} : t \in T\}$ ,  $r \in R$ .

Вероятность  $p(u, r)$  того, что клиент  $u$  выберет ресурс  $r$ , выписывается по формуле полной вероятности, причём сразу двумя способами:

$$p(u, r) = \sum_{t \in T} p_u p_{tu} q(r|t, u) = \quad (1)$$

$$= \sum_{t \in T} q_r q_{tr} p(u|t, r), \quad (2)$$

где  $p_u = p(u)$  и  $q_r = q(r)$  — априорные вероятности появления клиента  $u$  и ресурса  $r$  в записи протокола. Эти вероятности легко оцениваются по протоколу  $D$  как соответствующие частоты. Апостериорные вероятности выписываются по формуле Байеса:

$$q(r|t, u) = q(r|t) = q_{tr}q_r / \sum_{r' \in R} q_{tr'}q_{r'};$$

$$p(u|t, r) = p(u|t) = p_{tu}p_u / \sum_{u' \in U} p_{tu'}p_{u'}.$$

Подстановка апостериорных вероятностей в (1) и (2) позволяет выразить вероятность  $p(u, r)$  через неизвестные (скрытые) профили  $\{p_{tu}\}$  и  $\{q_{tr}\}$ . Предположим, что протокол  $D$  охватывает такой промежуток времени, в течение которого все рассматриваемые вероятности остаются неизменными. Тогда для восстановления скрытых профилей можно применить принцип максимума правдоподобия:

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

где максимум берётся по всем профилям  $\{p_{tu}\}$  и  $\{q_{tr}\}$ , удовлетворяющим ограничениям-равенствам  $\sum_{t \in T} p_{tu} = 1$  для всех  $u \in U$  и  $\sum_{t \in T} q_{tr} = 1$  для всех  $r \in R$ . Для решения данной оптимизационной задачи предлагается алгоритм, в котором чередуются два шага:

- оптимизация профиля  $\{p_{tu}\}$  при фиксированном  $\{q_{tr}\}$ ;
- оптимизация профиля  $\{q_{tr}\}$  при фиксированном  $\{p_{tu}\}$ .

На каждом шаге для оптимизации профиля выполняется несколько итераций EM-алгоритма. На каждой итерации возникает более простая задача максимума правдоподобия, которая решается аналитически.

Скрытыми переменными в EM-алгоритме являются апостериорные вероятности того, что клиент  $u$ , выбирая ресурс  $r$ , удовлетворяет свой интерес  $t$ . Эти оценки крайне важны для многих приложений.

Главной особенностью алгоритма является его «симметричность»: разложения по клиентам (1) и по ресурсам (2) наравне используются в итерациях, благодаря чему и достигается согласованность профилей.

Начальное приближение задаётся либо случайным образом, либо исходя из априорной информации. В частности, во многих приложениях ресурсы и/или клиенты изначально классифицированы по набору тем  $T$ . Это позволяет устранить т. н. «проблему холодного старта» — когда надо принимать решения относительно ресурса или клиента, для которого в  $D$  не зафиксировано ни одной записи. В таких случаях вычислить искомым профиль невозможно, но его можно заменить априорным профилем, имеющим ту же самую структуру.

Эксперименты на модельных данных ( $|R| = 200$ ,  $|U| = 1000$ ,  $|T| = 10$ ,  $\ell = 50\,000$ ), в которых истинные профили были известны изначально,

и точность их восстановления оценивалась в среднеквадратичном, показали, что 3–4 итераций на внешнем цикле и 4–5 ЕМ-итераций на внутренних циклах вполне достаточно для восстановления, причём дальнейшее увеличение числа итераций может даже немного ухудшить точность.

### Оценивание сходства клиентов и ресурсов

Существует масса различных способов определить функции расстояния между клиентами  $\rho(u, u')$  и между ресурсами  $\rho(r, r')$ : через корреляцию [3], через вероятность случайного совместного выбора [4], через ассоциативные правила, и др [5]. Имея профили, расстояние естественно определить как евклидову метрику в пространстве профилей.

Был проведён эксперимент на протоколах поисковой машины Яндекс. В данной задаче клиентами являются пользователи, делающие поисковые запросы; ресурсами являются документы, выдаваемые в результате поиска; действием пользователя считается переход по гиперссылке. Протокол охватывал семь дней регулярной работы Яндекса в 2005 году;  $|R| = 396$ ,  $|U| = 14606$ ,  $|T| = 20$ ,  $\ell = 1.3 \cdot 10^5$ . Множество ресурсов  $R$  было разбито по тематике на 8 классов. Строились две метрики на  $R$ : по вероятности совместных выборов [4] и по профилям ресурсов. Затем обе метрики использовались для классификации ресурсов методом  $k$  ближайших соседей. Число  $k$  настраивалось по скользящему контролю для каждой метрики отдельно. Доля ошибок оказалась равной 22% при  $k = 12$  для первой метрики и 11% при  $k = 15$  для второй метрики.

Таким образом, предложенный подход агрегирует информацию о пользователях и ресурсах в виде хорошо интерпретируемых профилей, и при этом строит более адекватные оценки сходства.

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

### Литература

- [1] Технология АКС — анализ клиентских сред. — ЗАО «Форецис». — 2005. <http://www.forecsys.ru/cea.php>.
- [2] *Marlin B.* Modeling user rating profiles for collaborative filtering // Neural Information Processing Systems (NIPS-16). — MIT Press, 2004.
- [3] *Sarwar B. M., Karypis G., Konstan J. A., Reidl J.* Item-based collaborative filtering recommendation algorithms // World Wide Web. — 2001. — Pp. 285–295.
- [4] *Воронцов К. В., Рудаков К. В., Лексин В. А., Ефимов А. Н.* Выявление и визуализация метрических структур на множествах пользователей и ресурсов Интернет // Искусств. Интеллект. — Донецк, 2006. — № 2 — С. 285–288.
- [5] *Symeonidis P., Nanopoulos A., Papadopoulos A., Manolopoulos Y.* Collaborative filtering: Fallacies and insights in measuring similarity // PKDD Workshop on Web Mining, Berlin, Germany. — 2006.