ПЕРЕОБУЧЕНИЕ В ВЕРОЯТНОСТНЫХ ЛАТЕНТНЫХ СЕМАНТИЧЕСКИХ МОДЕЛЯХ[1]

В. А. Лексин[2], К. В. Воронцов[3]

[2] МФТИ, Москва, vleksin@mail.ru
[3] ВЦ РАН, Москва, voron@ccas.ru

Предлагается алгоритм выявления скрытых интересов клиентов по наблюдаемому протоколу их действий, например, посещений сайтов. Алгоритм сочетает в себе идеи анализа клиентских сред и вероятностного латентного семантического анализа. Для оптимизации параметров алгоритма и объективного сравнения его с другими алгоритмами вводится критерий качества, основанный на классификации заранее размеченного множества сайтов. Эксперименты показывают, что качество имеет оптимум по основным параметрам алгоритма, и что попытка чрезмерно точной оптимизации может приводить к переобучению.

Автоматическое выявление потребностей и интересов клиентов по данным об их поведении (покупок, посещений, запросов, и т. д.) является актуальной задачей для многих сфер бизнеса, ориентированных на клиентов. Например, для персонализации предложений в рекомендующих системах, сегментации клиентской базы в маркетинговых исследованиях, поиска единомышленников в социальных сетях, выявления совместных продаж, и т. д.

Исходные данные представляют собой последовательность записей «клиент *u* выбрал ресурс *r*». Для успешного решения упомянутых задач необходимо адекватно оценивать сходство клиентов и ресурсов. Анализ клиентских сред (АКС) [1,7,9] основан на принципе *согласованного сходства*: «ресурсы схожи, если ими пользуются схожие клиенты; в то же время, клиенты схожи, если они пользуются схожими ресурсами». Простейшие методы анализа поведения пользователей Интернет (web usage mining, WUM) [2] и коллаборативной фильтрации (collaborative filtering, CF) [3] опираются либо только на сходство клиентов (user-based CF), либо только на сходство ресурсов (item-based CF). Несимметричность анализа относительно двойственных сущностей — клиентов и ресурсов — порождает ряд проблем и ограничивает применимость этих методов. Этого недостатка лишён латентный семантический анализ (latent semantic analysis, LSA), выявляющий взаимосвязи между клиентами и ресурсами через скрытые параметры [4]. Вероятностные модели (probabilistic LSA или PLSA) [5,6] имеют более глубокие статистические обоснования по сравнению с обычным LSA. Основная идея PLSA заключается в восстановлении скрытых параметров (профилей), характеризующих каждого клиента и каждый ресурс; обычно для этого применяется EM-алгоритм.

В данной работе предлагается подход, сочетающий принцип согласованного сходства АКС и оценивание скрытых профилей по PLSA. Это приводит к симметричному варианту EM-алгоритма с дополнительным внешним циклом итераций. Восстановленные профили легко сравнивать, что позволяет применять метод *k* ближайших соседей для классификации ресурсов и вводить объективные критерии качества для сравнения различных методов CF и WUM. В экспериментах исследуется зависимость качества классификации от длины скрытых профилей и числа итераций на внутреннем и внешнем цикле алгоритма. Оказывается, что качество имеет оптимум по всем этим параметрам, то есть попытка чрезмерно точной настройки на выборку может приводить к переобучению.

[1] Работа выполнена при поддержке РФФИ, проекты 07-01-12076-офи и 08-07-00422.

Восстановление скрытых профилей симметризованным EM-алгоритмом

Пусть заданы множество клиентов 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 клиентов $\rho_U(u,u')$ и ресурсов $\rho_R(r,r')$.

Допустим, что каждый клиент интересуется некоторым набором тем. Множество всех тем обозначим через T.

Профилем клиента $u \in U$ назовем вектор условных вероятностей $p_{tu} = p(t \mid u)$ того, что данный клиент u интересуется темой $t \in T$, причём $\sum_{t \in T} p_{tu} = 1$.

Аналогично, *профилем ресурса* $r \in R$ назовем вектор условных вероятностей $q_{tr} = q(t \mid r)$ того, что данный ресурс r удовлетворяет теме $t \in T$, причём $\sum_{t \in T} q_{tr} = 1$.

Требуется по протоколу D найти скрытые профили клиентов $\{p_{tu}, t \in T\}, u \in U$ и ресурсов $\{q_{tr}, t \in T\}, r \in R$.

Представим вероятность выбора клиентом u ресурса r двумя различными способами:

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

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

где p_u и q_r — априорные вероятности появления клиента u и ресурса r в записи протокола. Апостериорные вероятности $q(r \mid t,u)$ и $p(u \mid t,r)$ выражаются через профили по формуле Байеса.

Для нахождения профилей применим принцип максимума правдоподобия:

$$\sum_{i=1}^{l} \ln p(u_i, r_i) \to \max, \quad (3)$$

где максимум берется по всем профилям $\{p_{tu}\}$, $\{q_{tr}\}$ при ограничениях нормировки $\sum_{t \in T} p_{tu} = 1$, $u \in U$, и $\sum_{t \in T} q_{tr} = 1$, $r \in R$.

Для решения оптимизационной задачи (3) используется итерационный алгоритм [7], на внешнем цикле которого выполняются два шага:

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

Каждый из двух шагов внешнего цикла реализуется с помощью EM-алгоритма и образует внутренний цикл, в свою очередь также состоящий из двух шагов, называемых «E-шаг» и «M-шаг». На E-шаге оцениваются скрытые переменные — апостериорные вероятности того, что клиент u, обращаясь к ресурсу r, интересуется темой t. Благодаря введению скрытых переменных функционал (3) распадается на сумму независимых функционалов по клиентам и ресурсам. Эти функционалы удаётся максимизировать аналитически. Благодаря этому на M-шаге оптимальные профили вычисляются по явным формулам и достаточно эффективно. Главная особенность алгоритма заключается в его симметричности относительно двух различных и в равной степени допустимых разложений (1) и (2). Оба разложения используются в итерационном процессе, что обеспечивает взаимную согласованность профилей клиентов и ресурсов.

Эксперименты и выводы

Алгоритм тестировался на реальных данных поисковой машины Яндекс и на модельных данных.

Данные поисковой машины представляли собой протокол переходов пользователей на документы (ресурсы), выданные в результатах поиска. Были выбраны 1024 наиболее посещаемых ресурсов и 7292 наиболее активных пользователей за 1 неделю работы поисковой системы. Строились профили длины $|T| = 12$. Смысл компонент профилей априори не задавался, тем не менее, по окончании итераций у сайтов схожей тематики выделялись одни и те же компоненты (таблица. 1), причём 10 из 12 компонент чётко интерпретируются.

[1] Работа выполнена при поддержке РФФИ, проекты 07-01-12076-офи и 08-07-00422.

Таблица 1. Примеры восстановления профилей сайтов

Музыка												
www.mp3real.ru	0	0.01	0.86	0	0.02	0.04	0.01	0	0.03	0	0.01	0.01
mp3.musicfind.ru	0	0	0.96	0	0	0	0	0	0	0.02	0	0.01
akkordi.ru	0	0.01	0.85	0.02	0.03	0.02	0.01	0	0.01	0.02	0.01	0.03
www.muzzone.com	0.01	0	0.94	0	0	0	0.02	0	0	0.01	0	0.02
mp3forum.ru	0.01	0.01	0.85	0.02	0	0.01	0.04	0.01	0.01	0.03	0	0.01
Сотовая связь												
mindmix.ru/mobile	0.01	0.83	0.02	0	0.01	0.01	0.04	0	0.01	0.05	0	0
www.sotoman.ru	0.01	0.78	0.01	0.02	0.04	0.01	0.04	0.02	0.01	0.03	0.01	0.02
www.mobyline.ru	0.02	0.74	0.02	0.01	0.02	0.01	0.03	0.03	0.07	0.02	0.02	0.01
www.eurotel.ru	0.01	0.87	0.04	0	0.01	0.01	0.01	0	0	0.01	0.02	0.03
www.sota1.ru	0.01	0.91	0.01	0.01	0.01	0	0.02	0	0	0.01	0.01	0
Игры												
gameguru.ru	0.01	0.01	0	0.01	0.02	0.03	0.77	0.01	0.02	0.09	0.01	0.02
www.gameland.ru	0.08	0.01	0.02	0.02	0	0	0.73	0.05	0.02	0.05	0.01	0
www.ag.ru	0	0.02	0.04	0.01	0.01	0.02	0.84	0.01	0	0.01	0.01	0.04
www.neogame.ru	0.02	0.01	0	0	0.04	0.01	0.81	0.04	0.01	0.04	0.01	0.02

Существует множество способов ввести метрики на клиентах $\rho_U(u,u')$ и ресурсах $\rho_R(r,r')$. Наиболее очевидный — средний квадрат отклонения между профилями. Лучшие результаты получались при предварительном обнулении в каждом профиле всех компонент кроме двух максимальных.

По полученной метрике методами многомерного шкалирования строилась плоская карта сходства сайтов (рис. 1). Сайты схожей тематики образуют на карте достаточно чётко выделяемые кластеры. В каждом кластере профили сайтов имеют, как правило, одни и те же максимальные компоненты (пример в таблице. 1).

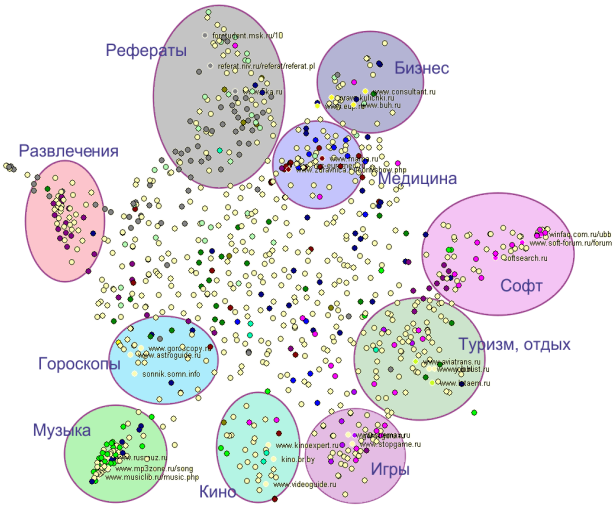

Рис. 1. Карта сходства.

Для сравнения качества профилей было размечено 396 сайтов на 12 классов. Критерий качества профилей определялся как доля размеченных сайтов, у которых максимум в профиле приходится на тот же элемент, что и в среднем по классу. На рис. 2 показан результат покоординатной оптимизации параметров алгоритма по этому критерию. По оси абсцисс отложены значения трёх параметров: число итераций внутреннего цикла EM-алгоритма, число итераций внешнего цикла, число тем $|T|$. По оси ординат отложены значения критерия. Лучшее качество достигалось при 8 итерациях на внешнем цикле, двух EM-итерациях на внутреннем цикле, и длине профиля 12, что совпало с числом классов. Дальнейшее увеличение числа итераций только ухудшает качество профиля. Это можно интерпретировать как переобучение при попытке избыточно точной настройки на конкретную выборку.

Для сравнения различных метрик вводился другой критерий — число ошибок классификации размеченных сайтов методом k ближайших соседей, при оптимальном k. Сравнивались три метрики на ресурсах — расстояние между профилями, корреляция посещений [8] и вероятность случайного совместного выбора [9]. Результаты, соответственно: 11%, 38% и 25% ошибок классификации. Таким образом, метрика, определяемая через профили, обладает существенно лучшим качеством (рис. 3).

[1] Работа выполнена при поддержке РФФИ, проекты 07-01-12076-офи и 08-07-00422.

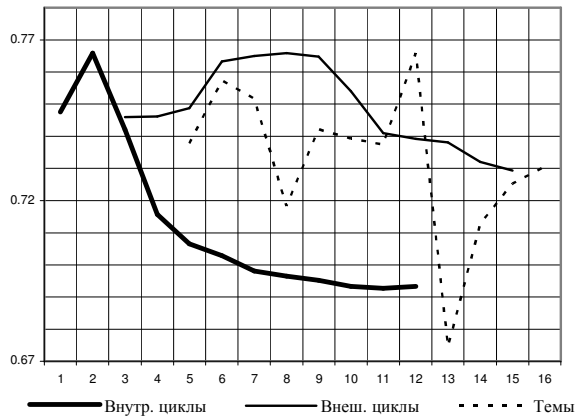

Рис. 2. Зависимость доли правильно восстановленных профилей от основных параметров алгоритма.

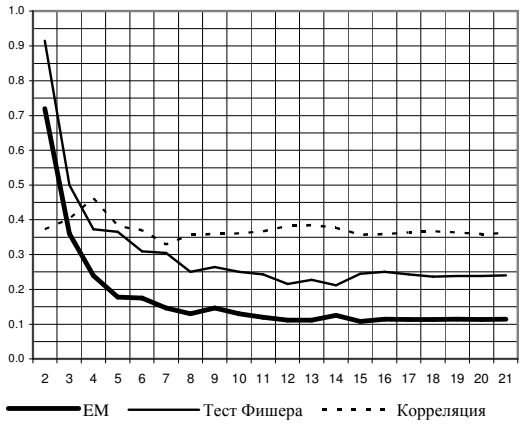

Рис. 3. Сравнение метрик. Зависимость доли ошибок классификации методом *k*NN от числа соседей *k*.

В эксперименте на модельных данных при $|R|=500$, $|U|=1000$, истинные профили задавались путём случайного выбора двух тем в каждом профиле. Выборка посещений генерировалась согласно вероятностной модели (1). Качество восстановления профилей оценивалось как средние по модулю отклонение от истинных профилей, при этом в восстановленных профилях выделялись два максимума, остальные компоненты обнулялись.

Оптимизация параметров на модельных данных дала оптимум при 6 итерациях на внешнем цикле и (снова) двух EM-итерациях на внутреннем цикле.

На модельных данных изучалась также расходимость алгоритма. Ставилась задача выяснить, при каком минимальном количестве тем и минимальной длине выборки алгоритм не расходится. Количество тем в исходных и восстанавливаемых профилях задавалось равным. В условиях данного эксперимента оказалось, что при количестве тем менее 10 или длине выборки менее 700 алгоритм расходится.

Работа выполнена при поддержке РФФИ, проекты 07-01-12076-офи и 08-07-00422.

Список литературы

1. Технология анализа клиентских сред (АКС) // ЗАО Форексис. — 2005. www.forecsys.ru/cea.php.
2. J. Fürnkranz. Web Mining // The Data Mining and Knowledge Discovery Handbook. 2005. Pp. 899–920.
3. J. Breese, D Heckerman, C Kadie. Empirical analysis of predictive algorithms for collaborative filtering // 14th annual conference on Uncertainty in Artificial Intelligence. — 1998. — P. 43–52.
4. M. Grčar. User Profiling: Collaborative Filtering // SIKDD 2004 at multiconference IS 12-15 Oct 2004, Ljubljana, Slovenia.
5. T. Hofmann. Latent Semantic Models for Collaborative Filtering // ACM Transactions on Information Systems, Vol. 22, No. 1, 2004, Pp. 89–115.
6. X. Jin, Y. Zhou, B. Mobasher. Web Usage Mining Based on Probabilistic Latent Semantic Analysis // Proc. of the 10th ACM SIGKDD international conference on Knowledge discovery and data mining. — 2004. — Pp. 197–205.
7. В. А. Лексин, К. В. Воронцов. Анализ клиентских сред: выявление скрытых профилей и оценивание сходства клиентов и ресурсов // ММРО-13, всеросс. конф. Математические методы распознавания образов. — М.: МАКС Пресс, 2007. — С. 488–491.
8. P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, J. Riedl. GroupLens: An Open Architecture for Collaborative Filtering of Netnews // In CSCW '94: Conference on Computer Supported Coorperative Work, Chapel Hill, ACM, P. 175–186.
9. К В. Воронцов, К. В. Рудаков, В. А. Лексин, А. Н. Ефимов // Выявление и визуализация метрических структур на множествах пользователей и ресурсов Интернет. // Искусственный Интеллект. — Донецк, 2006. — С. 285–288.

[1] Работа выполнена при поддержке РФФИ, проекты 07-01-12076-офи и 08-07-00422.