Двухступенчатые модели и проблема переобучения в латентном семантическом анализе*

В. А. Лексин, vleksin@gmail.com

Moscow Institute of Physics and Technology, 141700, 9, Institutskii per., Dolgoprudny, Moscow Region, Russia

Аннотация

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. В частности, такие задачи приходится решать с целью персонализации предложений в рекомендующих системах, сегментации клиентской базы в маркетинговых исследованиях, поиска единомышленников в социальных сетях 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] 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], такие как метод корреляций Пирсона или метод линейного сходства, опираются либо только на сходство клиентов (user-based CF), либо только на сходство ресурсов (item-based CF). Необходимость хранить все исходные данные, а также несимметричность анализа относительно двойственных сущностей — клиентов и ресурсов, ограничивает применимость этих методов. Этого недостатка лишён латентный семантический анализ (latent semantic analysis, LSA) [4, 5, 6], позволяющий выявлять латентные характеристики (профили) клиентов и ресурсов и заменять исходные данные более сжатым описанием. Как правило, для этого используются различные виды матричных разложений. Вероятностные модели (probabilistic LSA или pLSA) [7] имеют более

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

глубокие статистические обоснования и позволяют строить хорошо интерпретируемые профили клиентов и ресурсов. Обычно для этого применяется EM-алгоритм. Более подробный обзор методов коллаборативной фильтрации даётся в разделе 1.

В данной работе предлагается подход, сочетающий принцип согласованного сходства из анализа клиентских сред и оценивание латентных профилей из вероятностного семантического анализа [8, 1]. Это приводит к симметризованному двухступенчатому варианту EM-алгоритма с двумя вложенными циклами итераций, в отличие от стандартного алгоритма pLSA, где скрытые переменные вычисляются в одном цикле итераций. В разделе 2 этот метод рассматривается детально и сравнивается с алгоритмами LSA и pLSA.

Восстановленные профили легко сравнивать, что позволяет применить метод k ближайших соседей для классификации ресурсов и ввести объективные критерии качества для сравнения различных методов коллаборативной фильтрации.

В разделе 3 будут приведены результаты экспериментов и сравнительный анализ трех методов коллаборативной фильтрации (pLSA, двухступенчатый pLSA и корреляционный метод на основе точного теста Фишера) на реальных данных поисковой машины и данных о покупках товаров в крупной мебельной компании.

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 этим параметрам, то есть попытка чрезмерно точной настройки на выборку может приводить к переобучению.

1 Задачи и методы коллаборативной фильтрации

Коллаборативная фильтрация, как направление исследований, стала активно развиваться с конца 80-х, когда многие компании столкнулись с проблемой эффективного использования огромных объёмов «сырых» данных о поведении клиентов для решения ряда актуальных бизнес-задач, таких как персонализация услуг и направленный маркетинг [9, 10]. В основе коллаборативной фильтрации лежит предположение, что схожие клиенты имеют схожие предпочтения при выборе ресурсов. Это означает, что если для данного клиента найдено множество схожих с ним клиентов, то, зная их предпочтения, можно предсказывать, какие ресурсы предпочёл бы данный клиент. Это позволят формировать индивидуальные направленные предложения клиентам в виде ранжированных списков ресурсов. Методы коллаборативной фильтрации не анализируют содержимое (контент) ресурсов, поэтому они применимы для анализа ресурсов различного типа, в широком спектре прикладных областей. Более того, коллаборативная фильтрация позволяет определять сходство ресурсов, которые не имеют схожего контента, но неявно связаны через группы клиентов (communities), которые ими воспользовались. Многие методы коллаборативной фильтрации позволяют также выявлять тематику интересов каждого клиента. Именно такие методы и рассматриваются в данной работе.

1.1 Задача коллаборативной фильтрации

Пусть U — множество клиентов (субъектов, пользователей — users) , R — множество ресурсов (объектов, предметов — items), Y — пространство описаний транзак-

ций (либо фактов предпочтения, либо оценок ресурса клиентом). Исходные данные представляют собой протокол транзакций (database of users' preferences) — последовательность l троек $D = (u_i, r_i, y_i)_{i=1}^{l} \subset U \times R \times Y$.

Обычно протокол транзакций агрегируется в матрицу кросс-табуляции $F = ||f_{ur}||_{U \times R}$, где $f_{ur} = \text{aggr}\big\{(u_i, r_i, y_i) \in D \;\big|\; u_i = u,\ r_i = r\big\}$. Конкретный вид функции агрегирования aggr зависит от предметной области и конкретной задачи.

Основными задачами коллаборативной фильтрации являются: прогнозирование незаполненных ячеек f_{ur}, оценивание функций сходства (similarity functions) $K(u, u')$, $K(r, r')$, $K(u, r)$ между клиентами и ресурсами, выявление содержательно интерпретируемых латентных характеристик (профилей) клиентов и ресурсов.

Информация в протоколе транзакций может накапливаться в виде неявных рейтингов (implicit rating), либо в виде явных рейтингов (explicit rating).

Неявный рейтинг получается путем мониторинга действий клиента. Например, если U — множество пользователей Интернет, R — множество ресурсов (сайтов, документов, новостей и т. п.), то протокол посещений пользователей агрегируется в матрицу кросс-табуляции $F = ||f_{ur}||$, где f_{ur} — количество посещений ресурса r пользователем u. Наиболее типичные постановки задач: выдать оценку ресурса r для пользователя u; выдать пользователю u ранжированный список рекомендуемых ресурсов; сгенерировать для ресурса r список схожих с ним ресурсов.

В случае явных рейтингов клиенты сами проставляют оценки выбранным ресурсам. Например, если U — клиенты Интернет-магазина, R — товары (книги, видео, музыка, и т. п.), то значением f_{ur} может быть оценка, которую клиент u выставил товару r. Как правило, оценки принимают дискретные значения, $Y = 1, 2, \ldots, z_{\max}$. Задачи персонализации в этом случае ставятся аналогичным образом.

Алгоритмы коллаборативной фильтрации принято делить на два больших класса: анамнестические (memory-based) и модельные (model-based).

1.2 Анамнестические алгоритмы

Анамнестические алгоритмы [9, 11, 12] основаны на хранении всей матрицы кросс-табуляции F и непосредственном поиске в ней схожих клиентов (строк) и ресурсов (столбцов). Значение неизвестного рейтинга $\hat{f}_{ur}$ для клиента u и ресурса r оценивается по совокупности рейтингов, проставленных ресурсу r другими клиентами, предпочтения которых наиболее схожи с клиентом u:

$$\hat{f}_{ur} = \underset{u' \in U_\alpha(u)}{\text{aggr}} f_{u'r},$$

где $U_\alpha(u) = \{u' \in U \mid K(u', u) > \alpha\}$ — множество клиентов, схожих с u. Функция сходства клиентов $K(u, u')$ принимает тем большие значения, чем более схожи предпочтения клиентов u и u'. Параметр α задаёт пороговое значение сходства.

Примером агрегирующей функции aggr является формула непараметрического (ядерного) сглаживания Надарайя-Ватсона со сглаживающим ядром $K(u, u')$:

$$\hat{f}_{ur} = \bar{f}_u + \frac{\sum\limits_{u' \in U_\alpha(u)} K(u, u')(f_{u'r} - \bar{f}_{u'})}{\sum\limits_{u' \in U_\alpha(u)} K(u, u')},$$

где $\bar{f}_u = \frac{1}{|R(u)|} \sum\limits_{r \in R(u)} f_{ur}$ — средняя оценка клиента u, $R(u)$ — множество ресурсов, выбранных клиентом u. Чем более схожи предпочтения клиентов u и u', тем большим будет вклад рейтинга $f_{u'r}$ в прогнозируемую величину $\hat{f}_{ur}$.

В коллаборативных рекомендующих системах используются разные подходы к оцениванию сходства между клиентами $K(u, u')$. Пусть $R(u, u')$ — множество ресурсов, выбранных обоими клиентами u и u'.

В корреляционном подходе [9, 12] сходство клиентов u и u' оценивается по коэффициенту корреляции Пирсона:

$$K(u,u') = \frac{\sum\limits_{r \in R(u,u')} (f_{ur} - \bar{f}_u)(f_{u'r} - \bar{f}_{u'})}{\sqrt{\sum\limits_{r \in R(u,u')} (f_{ur} - \bar{f}_u)^2 \sum\limits_{r \in R(u,u')} (f_{u'r} - \bar{f}_{u'})^2}}.$$

При подходе, использующем метод линейного сходства [12, 11], клиенты u и u' представляются как векторы m-мерного пространства, $m = |R(u, u')|$, и сходство оценивается как косинус угла между этими векторами:

$$K(u,u') = \frac{\sum\limits_{r \in R(u,u')} f_{ur} f_{u'r}}{\sqrt{\sum\limits_{r \in R(u,u')} f_{ur}^2 \sum\limits_{r \in R(u,u')} f_{u'r}^2}}.$$

Ещё одна функция сходства, применяемая к бинарным матрицам кросс-табуляции, основывается на точном тесте Фишера [13]. Мы рассмотрим эту функцию сходства относительно ресурсов, а не относительно клиентов, поскольку именно в такой форме она будет использована в экспериментальной части данной работы. Вообще, большинство функций сходства, используемых в коллаборативной фильтрации, могут определяться относительно как клиентов, так и ресурсов. Будем оценивать сходство ресурсов r и r' путём проверки статистической гипотезы о независимости предпочтений клиентов, выбравших хотя бы один из двух ресурсов r и r'. Пусть $U(r)$ и $U(r')$ — множества клиентов, которые предпочли либо только ресурс r, либо только ресурс r' соответственно, $U(r, r')$ — множество клиентов, выбравших оба ресурса. Если мощность множества $U(r, r')$ настолько велика, что вероятность чисто случайного совместного выбора обоих ресурсов

$$P(r,r') = \mathsf{P}\big(|U(r_i, r_j)| = x\big) = \frac{C_{|U(r)|}^{x} C_{|U|-|U(r)|}^{|U(r')|-x}}{C_{|U|}^{|U(r')|}}$$

меньше заданного уровня значимости α, то можно полагать, что наблюдаемые данные противоречат гипотезе независимости, следовательно, между посещениями данной пары ресурсов имеется закономерная связь. Чем меньше значение $P(r, r')$, тем более схожи ресурсы. Функция сходства определяется как некоторая убывающая функция вероятности, например, $K(r, r') = -\log P(r, r')$.

1.3 Модельные алгоритмы

В отличие от анамнестических методов, модельные алгоритмы не хранят в памяти ни исходный протокол, ни матрицу кросс-табуляции F. Вместо этого для каж-

дого клиента и каждого ресурса формируется векторное описание (профиль). Функции сходства клиентов и ресурсов реализуются путём непосредственного сравнения этих профилей. В некоторых моделях компоненты профилей имеют содержательную интерпретацию. В частности, они могут соответствовать типам или тематикам ресурсов, интересам или социально-демографическим характеристикам клиентов. Есть и такие модели, в которых профили не имеют содержательной интерпретации.

В моделях латентного семантического анализа, основанных на матричной факторизации [4, 6, 5], и вероятностного латентного семантического анализа (модель аспектов) [14] профили характеризуют интересы клиентов и тематику ресурсов. Благодаря возможности содержательной интерпретации профилей и наличию эффективных численных методов эти модели стали достаточно популярны [12, 3, 11].

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

Профилем клиента $u \in U$ назовем вектор условных вероятностей $p_{tu} = p(t|u)$ того, что данный клиент u интересуется темой $t \in T$. Профиль должен удовлетворять условию нормировки $\sum_{t \in T} p_{tu} = 1$.

Профилем ресурса $r \in R$ назовем вектор условных вероятностей $q_{tr} = q(t|r)$ того, что данный ресурс r соответствует теме $t \in T$. Аналогично, $\sum_{t \in T} q_{tr} = 1$.

Здесь и далее все вероятности, относящиеся к клиентам, обозначаются буквой p, а вероятности, относящиеся к ресурсам — буквой q.

1.3.1 Вероятностный латентный семантический анализ

Вероятностный латентный семантический анализ (pLSA) основан на выявлении латентных характеристик (профилей) для каждого клиента и каждого ресурса.

Рассмотрим вероятностную модель предпочтения ресурса r клиентом u:

$$p(u, r) = \sum_{t \in T} p(t) p(u|t) q(r|t), \tag{1}$$

где $p(t)$ — априорная вероятность, характеризующая «популярность» темы $t \in T$, $p(u|t)$ — апостериорное распределение клиентов по каждой теме t, $q(r|t)$ — апостериорное распределение ресурсов по каждой теме t.

Предполагается, что число тем $|T|$ существенно меньше числа клиентов $|U|$, числа ресурсов $|R|$ и длины протокола l. При этих условиях можно воспользоваться принципом максимума правдоподобия, чтобы найти неизвестные в данной модели параметры $p(t)$, $p(u|t)$ и $q(r|t)$ по наблюдаемой матрице кросс-табуляции $F = \|f_{ur}\|_{U \times R}$, где f_{ur} — число раз, когда клиент u выбирал ресурс r:

$$L = \ln \prod_{u \in U} \prod_{r \in R} p(u, r)^{f_{ur}} = \sum_{u \in U} \sum_{r \in R} f_{ur} \ln p(u, r) \to \max_{p(t),\, p(u|t),\, q(r|t)}.$$

Для поиска оценки максимального правдоподобия используется алгоритм Expectation-Maximization (EM) [15]. Идея этого алгоритма заключается в следующем. Для оценки неизвестных параметров $p(t)$, $p(u|t)$ и $q(r|t)$ вводится вспомогательный вектор скрытых переменных $p(t|u, r)$, который можно интерпретировать как вероятность того, что клиент u, выбирая ресурс r, удовлетворял свой интерес t.

Вектор скрытых переменных обладает двумя свойствами. С одной стороны, он может быть легко вычислен, если известны значения параметров $p(t)$, $p(u|t)$ и $q(r|t)$. С другой стороны, решение задачи максимизации правдоподобия сильно упрощается, если известны значения скрытых переменных. EM-алгоритм состоит из итерационного повторения двух шагов.

На E-шаге (Expectation) ожидаемое значение вектора скрытых переменных $p(t|u,r)$ вычисляется по формуле Байеса на основе текущих значений неизвестных параметров:

$$p(t|u,r) = \frac{p(t)p(u|t)q(r|t)}{\sum\limits_{t' \in T} p(t')p(u|t')q(r|t')}, \quad u \in U,\ r \in R,\ t \in T.$$

На M-шаге (Maximization) решается задача максимизации правдоподобия и находится следующее приближение неизвестных параметров $p(t)$, $p(u|t)$ и $q(r|t)$. Эту задачу удаётся решить аналитически, воспользовавшись значениями скрытых переменных, найденных на E-шаге. Решение записывается следующим образом:

$$p(t) = \frac{\sum\limits_{u \in U} \sum\limits_{r \in R} f_{ur} p(t|u,r)}{\sum\limits_{u \in U} \sum\limits_{r \in R} f_{ur}}, \quad t \in T;$$

$$q(r|t) = \frac{\sum\limits_{u \in U} f_{ur} p(t|u,r)}{\sum\limits_{u \in U} \sum\limits_{r' \in R} f_{ur'} p(t|u,r')}, \quad r \in R,\ t \in T;$$

$$p(u|t) = \frac{\sum\limits_{r \in R} f_{ur} p(t|u,r)}{\sum\limits_{u' \in U} \sum\limits_{r \in R} f_{u'r} p(t|u',r)}, \quad u \in U,\ t \in T.$$

Далее происходит возврат к E-шагу при новых значениях параметров $p(t)$, $p(u|t)$ и $q(r|t)$. Итерации продолжаются, пока не произойдёт стабилизация значений параметров и/или значения правдоподобия. Начальные приближения для $p(t)$, $p(u|t)$ и $q(r|t)$ инициализируются случайными или равномерными распределениями.

Найденные параметры $p(u|t)$ и $q(r|t)$ являются условными распределениями клиентов и ресурсов относительно каждой темы t. Однако искомые профили клиентов и ресурсов должны иметь вид условных распределений тем $p(t|u)$ и $q(t|r)$. Для их вычисления достаточно применить формулу Байеса:

$$p_{tu} = p(t|u) = \frac{p(t)p(u|t)}{\sum\limits_{t' \in T} p(t')p(u|t')}, \quad u \in U,\ t \in T;$$

$$q_{tr} = q(t|r) = \frac{p(t)q(r|t)}{\sum\limits_{t' \in T} p(t')q(r|t')}, \quad r \in R,\ t \in T.$$

1.3.2 Латентный семантический анализ (матричные разложения)

Латентный семантический анализ (latent semantic analysis, LSA) [4, 6] основан на матричных разложениях. Пусть T — множество тем (интересов). Обычно предполагается, что $|T| \ll |U|$ и $|T| \ll |R|$. p_{tu} — неизвестный профиль клиента u, q_{tr} — неиз-

вестный профиль объекта r. Введем матрицы $P = (p_{tu})_{|T|\times|U|}$, $Q = (q_{tr})_{|T|\times|R|}$ и диагональную матрицу Λ размера $|T| \times |T|$, элементы которой $\lambda_{kk} = \lambda(t_k)$, $k = 1, \ldots, |T|$ содержат собственные числа матрицы F. Нашей задачей является найти разложение $f_{ur} = \sum\limits_{t\in T} \lambda_t p_{tu} q_{tr}$, что в матричном виде будет выглядеть как

$$\hat{F} = P\Lambda Q'. \tag{2}$$

Согласно теореме о сингулярном разложении матриц (singular value decomposition, SVD) [7], любая матрица, например, матрица кросс-табуляции $F_{|U|\times|R|}$ может быть представлена в виде

$$F = P_0 \Lambda_0 Q_0', \tag{3}$$

где P_0 — матрица размера $|U| \times K$ такая, что $P_0 P_0' = I$; Q_0 — матрица размера $K \times |R|$ такая, что $Q_0 Q_0' = I$; Λ_0 — диагональная матрица размера $K \times K$, содержащая собственные числа F; K — ранг матрицы F, $K \leq \min(|U|, |R|)$.

В общем случае P_0, Λ_0 и Q_0 — матрицы полного ранга. Особенностью сингулярного разложения является то, что оно позволяет получить приближенное значение F, используя матрицы меньшей размерности. Оставим наибольшие $|T|$ собственных чисел в Λ_0, а остальные обнулим. При этом результирующая матрица $\hat{F}$ будет лишь приблизительно равна F и иметь ранг $|T|$. Можно показать, что таким образом полученная матрица $\hat{F}$ из всех матриц ранга, не превышающего T, наиболее близка к F по среднеквадратичному отклонению. Поскольку в матрицу Λ_0 были добавлены нули, то можно выбросить нулевые столбцы и строки из Λ_0, а также соответствующие столбцы и строки в P_0 и Q_0, чтобы получить матрицы P и Q и Λ соответственно. В результате получим:

$$F \approx \hat{F} = P\Lambda Q'.$$

Тем самым мы получаем искомое разложение (2). Фактически, в матрице $\hat{F}$ мы получаем значения неизвестных рейтингов для всех ресурсов и клиентов. Если мы хотим оценить сходство клиентов или ресурсов между собой, то можем воспользоваться следующими моделями: $\hat{F}'\hat{F} = Q\Lambda^2 Q'$ — матрица сходства ресурсов, $\hat{F}\hat{F}' = P\Lambda^2 P'$ — матрица сходства клиентов.

Недостатком сингулярного разложения является то, что профили p_{tu} и q_{tr} тяжело поддаются содержательной интерпретации.

Лучше интепретируется неотрицательное матричное разложение (non-negative matrix factorization, NNMF) [16], где $p_{tu} \geqslant 0$ и $q_{tr} \geqslant 0$. В этом случае элементы профиля p_{tu} можно интерпретировать как апостериорные вероятности $p(u|t)$, q_{tr} как апостериорные вероятности $q(r|t)$, а элементы матрицы Λ как $\lambda_t = p(t)$. В этом случае можно провести прямую аналогию формулы (2) с (1), то есть метода NNMF с вероятностным латентным семантическим анализом (pLSA).

На практике явное решение задачи на собственные числа требуют больших затрат машинного времени, поэтому в [17] предложен итерационный процесс, который позволяют оценить P, Q и Λ без прямого вычисления собственных чисел.

2 Двухступенчатая (симметризованная) вероятностная латентная модель

В данной работе предлагается двухступенчатая (симметризованная) модель вероятностного латентного семантического анализа, в которой профили клиентов и ресурсов уточняются попеременно. Эксперименты показывают, что при этом повышается точность восстановления профилей и скорость сходимости алгоритма.

Пусть $F = \|f_{ur}\|_{U\times R}$ — матрица исходных данных, где f_{ur} — число раз, когда клиент $u \in U$ выбирал ресурс $r \in R$.

Вместо вероятностной модели предпочтения (1) запишем другое, формально ей эквивалентное, выражение для вероятности выбора ресурса r клиентом u:

$$p(u,r) = \sum_{t\in T} p(u)p(t|u)q(r|t,u), \tag{4}$$

где $p(u) = p_u$ — априорная вероятность того, что очередной выбор будет сделан клиентом u, $p(t|u) = p_{tu}$ — вероятность того, что клиент u интересуется темой t, $q(r|t,u) = q(r|t)$ — апостериорная вероятность того, что будет выбран ресурс r при условии, что этот выбор продиктован интересом к теме t. Гипотеза, что апостериорная вероятность $q(r|t,u)$ не зависит от клиента u, является необходимым в данной модели упрощающим предположением. Априорные вероятности и профили должны удовлетворять условиям нормировки $\sum\limits_{u\in U} p_u = 1$ и $\sum\limits_{t\in T} p_{tu} = 1$ для всех $u \in U$.

Апостериорную вероятность $q(r|t)$ выразим по формуле Байеса через априорные вероятности $q(r) = q_r$ и профили ресурсов $q(t|r) = q_{tr}$:

$$q(r|t) = \frac{q(t|r)q(r)}{\sum\limits_{r'\in R} q(t|r')q(r')} = \frac{q_{tr}q_r}{\sum\limits_{r'\in R} q_{tr'}q_{r'}}.$$

Подставим это выражение в формулу для $p(u,r)$:

$$p(u,r) = \sum_{t\in T} p_u p_{tu} \frac{q_r q_{tr}}{\sum\limits_{r'\in R} q_{tr'}q_{r'}}.$$

Априорные вероятности p_u и q_r легко оценить эмпирически как долю транзакций, в которых, соответственно, выбор делался клиентом u или был выбран ресурс r:

$$p_u = \frac{1}{l}\sum_{r\in R} f_{ur}; \qquad q_r = \frac{1}{l}\sum_{u\in U} f_{ur}; \qquad l = \sum_{u\in U}\sum_{r\in R} f_{ur}.$$

Таким образом, вероятность $p(u,r)$ выражается через известные априорные вероятности p_u, q_r и неизвестные профили клиентов p_{tu} и ресурсов q_{tr}. Заметим, что $p(u,r)$ линейно зависит от профилей клиентов p_{tu} и довольно сложным образом — от профилей ресурсов q_{tr}. Поэтому допустим, что профили ресурсов q_{tr} уже известны и фиксированы. Для нахождения профилей клиентов $\{p_{tu} : u \in U, t \in T\}$ будем решать задачу максимизации правдоподобия при $|U|$ ограничениях-равенствах:

$$L\big(\{p_{tu}\}\big) = \sum_{u\in U}\sum_{r\in R} f_{ur} \ln p(u,r) \to \max_{\{p_{tu}\}};$$
$$\sum_{t\in T} p_{tu} = 1, \quad u \in U.$$

Есть ещё ограничения-неравенства $p_{tu} \geq 0$, однако мы их пока исключим из рассмотрения, а позже убедимся, что они с гарантией выполняются.

Запишем Лагранжиан оптимизационной задачи:

$$L\big(\{p_{tu}\},\{\lambda_u\}\big) = \sum_{u\in U}\sum_{r\in R} f_{ur}\ln\Big(p_u\sum_{t\in T} p_{tu}q(r|t)\Big) - \sum_{u\in U}\lambda_u\Big(\sum_{t\in T} p_{tu} - 1\Big).$$

Продифференцируем Лагранжиан по p_{tu} и приравняем нулю производную:

$$\frac{\partial L}{\partial p_{tu}} = \sum_{r\in R} f_{ur}\frac{1}{p_{tu}}\frac{p_{tu}q(r|t)}{\sum\limits_{t'\in T} p_{t'u}q(r|t')} - \lambda_u = 0, \quad t\in T,\ u\in U. \tag{5}$$

Введем вспомогательные *скрытые переменные* $H_{tr}(u)$:

$$H_{tr}(u) = \frac{p_{tu}q(r|t)}{\sum\limits_{t'\in T} p_{t'u}q(r|t')}, \quad t\in T,\ r\in R,\ u\in U.$$

Заметим, что, согласно формуле Байеса, $H_{tr}(u) = H(t|r,u)$ есть апостериорная вероятность темы t для заданной пары (u,r). Другими словами, это вероятность того, что причиной выбора ресурса r клиентом u был интерес t. Очевидно, для любой пары $(u,r)\in U\times R$ выполняется условие нормировки $\sum\limits_{t\in T} H_{tr}(u) = 1$.

Допустим, что значения скрытых переменных $H_{tr}(u)$ известны и фиксированы. Умножим обе части равенства (5) на p_{tu} и просуммируем по t:

$$\sum_{t\in T}\sum_{r\in R} f_{ur}H_{tr}(u) = \lambda_u\sum_{t\in T} p_{tu}, \quad u\in U.$$

Переставляя знаки суммирования и учитывая условия нормировки $\sum\limits_{t\in T} p_{tu} = 1$ и $\sum\limits_{t\in T} H_{tr}(u) = 1$, получим $\lambda_u = \sum\limits_{r\in R} f_{ur}$. Подставляя λ_u обратно в (5), имеем:

$$p_{tu} = \frac{\sum\limits_{r\in R} f_{ur}H_{tr}(u)}{\sum\limits_{r\in R} f_{ur}}.$$

Поскольку значения скрытых переменных зависят от p_{tu}, для решения задачи максимизации правдоподобия приходится применять итерационный процесс, представляющий собой вариант EM-алгоритма. Каждая итерация состоит из двух шагов. На E-шаге фиксируются профили p_{tu} и согласно формуле Байеса вычисляются скрытые переменные $H_{tr}(u)$. На M-шаге фиксируются скрытые переменные $H_{tr}(u)$ и вычисляются профили p_{tu}. В качестве начального приближения можно взять равномерное распределение $p_{tu} = |T|^{-1}$. Заметим, что при любом неотрицательном начальном приближении гарантируется неотрицательность скрытых переменных и профилей p_{tu} на всех последующих итерациях.

Профили ресурсов q_{tr} до сих пор оставались фиксированными. Задача оптимизации профилей ресурсов q_{tr} при фиксированных профилях клиентов p_{tu} ставится и решается «симметричным образом», если буквально заменить во всех формулах

Алгоритм 2.1. Двухступенчатый симметризованный EM-алгоритм.

Вход:

матрица кросс-табуляции $F = \|f_{ur}\|_{U \times R}$;
число тем $|T|$;
число итераций на внешнем цикле I_{pq};
число итераций на внутреннем цикле I_{EM};

Выход:

p_{tu} — профили клиентов, q_{tr} — профили ресурсов;

оценки априорных вероятностей:
$p_u := \frac{1}{l} \sum_{r \in R} f_{ur}$; $q_r := \frac{1}{l} \sum_{u \in U} f_{ur}$ для всех $u \in U$, $r \in R$;
начальное приближение профилей:
$p_{tu} := |T|^{-1}$; $q_{tr} := |T|^{-1}$ для всех $u \in U$, $r \in R$, $t \in T$;
повторить I_{pq} раз внешний цикл итераций:
  $q(r|t) := \dfrac{q_{tr} q_r}{\sum_{r' \in R} q_{tr'} q_{r'}}$ для всех $r \in R$, $t \in T$;
  повторить I_{EM} раз внутренний цикл итераций:
    E-шаг: $H_{tr}(u) := \dfrac{p_{tu} q(r|t)}{\sum_{t' \in T} p_{t'u} q(r|t')}$ для всех $t \in T$, $u \in U$, $r \in R$;
    M-шаг: $p_{tu} := \dfrac{\sum_{r \in R} f_{ur} H_{tr}(u)}{\sum_{r \in R} f_{ur}}$ для всех $u \in U$, $t \in T$;
  $p(u|t) := \dfrac{p_{tu} p_u}{\sum_{u' \in U} p_{tu'} p_{u'}}$ для всех $u \in U$, $t \in T$;
  повторить I_{EM} раз внутренний цикл итераций:
    E-шаг: $H_{tu}(r) := \dfrac{q_{tr} p(u|t)}{\sum_{t' \in T} q_{t'r} p(u|t')}$ для всех $t \in T$, $u \in U$, $r \in R$;
    M-шаг: $q_{tr} := \dfrac{\sum_{u \in U} f_{ur} H_{tu}(r)}{\sum_{u \in U} f_{ur}}$ для всех $r \in R$, $t \in T$;

$u \leftrightarrow r$ и $q \leftrightarrow p$. Опуская дальнейшие выкладки, запишем только исходную вероятностную модель, которая формально эквивалентна моделям (1) и (4):

$$p(u, r) = \sum_{t \in T} q(r) q(t|r) p(u|t, r), \tag{6}$$

где $p(u|t, r) = p(u|t)$ — апостериорная вероятность того, что выбор будет сделан клиентом u при условии, что этот выбор продиктован интересом к теме t.

Основная идея двухступенчатого (симметризованного) EM-алгоритма заключается в организации двух вложенных циклов итераций. На внешнем цикле попеременно решаются две задачи: сначала профили клиентов p_{tu} оптимизируются при фиксированных профилях ресурсов q_{tr}, затем, наоборот, профили q_{tr} оптимизируются при фиксированных p_{tu}. Каждая оптимизация профилей реализуется внутренним циклом EM-алгоритма, в котором попеременно вычисляются то скрытые переменные $H(t|r, u)$, то очередное приближение профилей. Более подробно реализация этой идеи показана в Алгоритме 2.1.

Заметим, что эквивалентность всех трёх моделей (1), (4) и (6) следует из опре-

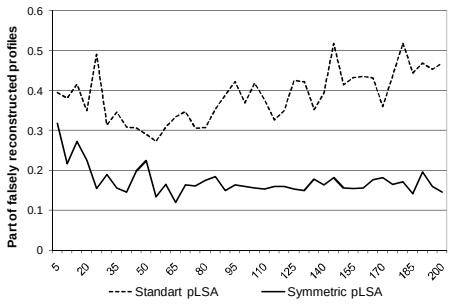

Рис. 1: Оптимизация количества тем.

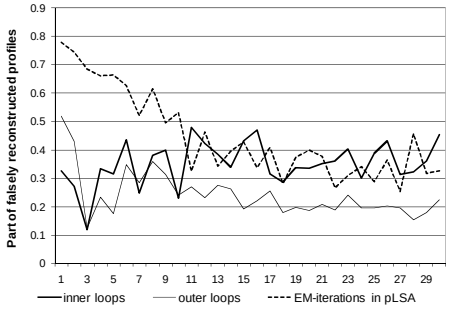

Рис. 2: Оптимизация количества итераций.

деления условной вероятности:

$$p(u)p(t|u) = p(u,t) = p(t)p(u|t), \quad u \in U,\ t \in T;$$
$$q(r)q(t|r) = q(r,t) = p(t)q(r|t), \quad r \in R,\ t \in T;$$

однако в стандартном алгоритме pLSA и в симметризованном алгоритме итерационный процесс организуется по-разному, имеет различные характеристики сходимости и приводит, вообще говоря, к разным результатам.

3 Эксперименты

Предложенный алгоритм тестировался и сравнивался со стандартными алгоритмами на данных поисковой машины «Яндекс», на данных о покупках товаров в крупной мебельной компании и на модельных данных.

3.1 Данные поисковой машины

Данные поисковой машины представляли собой протокол переходов пользователей (клиентов) на сайты или документы (ресурсы), выданные в результатах поиска. Лог-файл размером 3.7 Гб содержал данные о 129 600 ресурсах, 14 606 пользователях и 207 696 посещениях за 1 неделю работы поисковой системы «Яндекс». Для эксперимента были выбраны 1 024 наиболее популярных сайтов (посещавшихся более 30 раз) и 1 902 наиболее активных пользователей (сделавших не менее 30 посещений).

В эксперименте сравнивались различные методы построения функций сходства между сайтами, в том числе методы, основанные на построении профилей сайтов.

Для оценивания качества функций сходства были размечены 400 сайтов на 12 классов. Критерий качества определялся как доля ошибок при классификации размеченных сайтов методом k ближайших соседей. Разметка сайтов использовалась только для оценивания качества и не учитывалась при построении функций

	Компоненты профиля											
Сайт	1	2	3	4	5	6	7	8	9	10	11	12
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
Рефераты, учебные ресурсы												
www.zachetka.ru	0	0	0	0.01	**0.16**	**0.56**	0	0	0.02	0.01	0.21	0
edu.mton.ru	0	0	0	0.01	**0.45**	**0.41**	0	0	0.01	0	0.1	0
forstudent.msk.ru	0	0	0.01	0.01	**0.39**	**0.44**	0.01	0.01	0.02	0	0.1	0
www.5ka.ru	0.01	0.01	0	0.02	**0.11**	**0.65**	0.01	0.01	0.02	0.01	0.14	0.01
school.edu.ru	0.01	0.06	0.01	0.05	**0.53**	**0.17**	0.01	0.02	0.03	0.01	0.1	0.01
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

Таблица 1: Примеры восстановления профилей сайтов.

сходства. Функция сходства профилей строилась на основе среднеквадратичного отклонения с предварительным занулением неинформативных компонент.

Для построения профилей исследовались два алгоритма: стандартный pLSA и двухступенчатый (симметризованный) алгоритм. Для симметризованного алгоритма производилась покоординатная оптимизация параметров $|T|$, I_{pq} и I_{EM} по описанному выше критерию. После построения профилей в них зануялись все компоненты кроме трёх максимальных и производилась нормировка.

На рисунке 1 показаны результаты оптимизации количества тем $|T|$ в обоих алгоритмах. Наилучшее качество достигалось при $|T| = 65$ для симметризованного алгоритма и $|T| = 55$ в стандартном pLSA.

В симметризованном алгоритме внутри каждого их двух внутренних циклов вычисляются скрытые вероятности $H(t|u, r)$, но двумя различными способами. На рисунке 4 вычислятся средний модуль отклонения этих вероятностей в зависимости от числа внешних итераций. Согласованность профилей проявляется в том, что эти веростности сходятся друг к другу.

На рисунке 2 показаны результаты оптимизации количества внутренних и внешних итераций для двухступенчатого алгоритма и количества EM-итераций для стандартного pLSA. Лучшее качество для двухступенчатого алгоритма достигалось при трёх итерациях на внешнем цикле, трёх EM-итерациях на внутреннем цикле, а для стандартного pLSA — при 27 EM-итерациях.

Дальнейшее увеличение числа тем $|T|$ или числа итераций может ухудшать качество профиля. Это можно интерпретировать как переобучение при попытке избы-

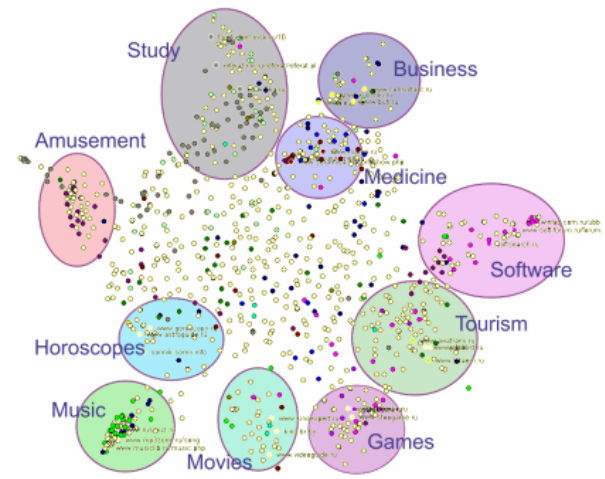

Рис. 3: Карта сходства ресурсов.

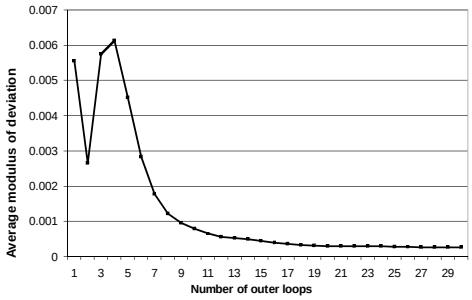

Рис. 4: Средний модуль отклонения скрытых вероятностей $H_{tr}(u)$ и $H_{tu}(r)$.

точно точной настройки на конкретную выборку.

Смысл компонент профилей априори не задавался, тем не менее, на данных Яндекса у сайтов схожей тематики выделялись одни и те же компоненты, причём некоторые компоненты чётко интерпретируются. В качестве примера в таблице 1 приведены профили длины $|T| = 12$ для некоторых сайтов. Этот пример показывает, что интерпретация компонент профилей может производиться уже после решения задачи, на основе анализа небольшой части ресурсов, тематика которых известна.

Существует множество способов ввести функции расстояния (метрики) на клиентах $\rho(u, u')$ и ресурсах $\rho(r, r')$. Наиболее очевидный — средний квадрат отклонения между профилями. По полученной метрике методами многомерного шкалирования строилась плоская карта сходства сайтов (рис. 3). Сайты схожей тематики образуют на карте достаточно чётко выделяемые кластеры. При этом в каждом кластере профили сайтов имеют, как правило, одни и те же максимальные компоненты (пример в таблице. 1). Наиболее чёткая кластеризация получалась при предварительном обнулении в каждом профиле всех компонент кроме трех максимальных.

3.2 Данные о покупке товаров

Данные о покупке товаров в мебельной компании представляли собой историю продаж за 3 года работы компании. Для анализа были отобраны 1 920 товаров, которые продавались более 100 раз и 1 328 постоянных клиентов, которые делали покупки более 30 раз. По выбранным клиентам и товарам анализировалась выборка из 112 256 фактов покупки товаров.

Для оценки качества по товарам использовалось разбиение 403 товаров на 12 категорий. Аналогично экспериментам на данных поисковой машины использовался метод k ближайших соседей (при $k = 5$) для оценки доли правильно классифицированных товаров. Оптимальное значение функционала (3% ошибок классификации) достигалось при количестве тем $|T|$=30, 4 внутренних и 4 внешних итерациях для симметризованного алгоритма. Также строились профили и метрика на множестве

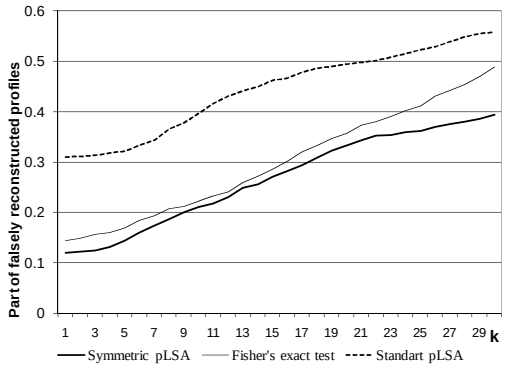

Рис. 5: Сравнение различных метрик по kNN на данных поисковой машины.

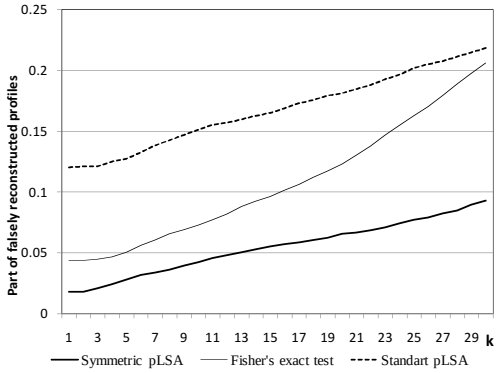

Рис. 6: Сравнение метрик на данных мебельной компании.

клиентов.

3.3 Сравнение различных метрик

Для сравнения различных алгоритмов коллаборативной фильтрации использовалась оценка качества классификации ресурсов методом k ближайших соседей. На рисунках 5 и 6 сравниваются три метрики на ресурсах: расстояние между профилями в симметризованном алгоритме, расстояние между профилями в стандартном pLSA и расстояние, основанное на точном тесте Фишера для данных поисковой машины и мебельной компании.

Таким образом, двухступенчатая аспектная модель быстрее сходится и обеспечивает более высокое качество функций сходства по сравнению со стандартным pLSA и корреляционной мерой сходства на основе точного теста Фишера.

3.4 Модельные данные

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 (4). 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 показала, что наилучшее качество восстановления профилей достигается при 6 итерациях на внешнем цикле и 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.

Список литературы

[1] *Лексин В. А., Воронцов К. В.* Анализ клиентских сред: выявление скрытых профилей и оценивание сходства клиентов и ресурсов // Математические методы распознавания образов-13. — М.: МАКС Пресс, 2007. — С. 488–491. www.mmro.ru .

[2] *Jin X., Zhou Y., Mobasher B.* Web usage mining based on probabilistic latent semantic analysis. — 2004. citeseer.comp.nus.edu.sg/701255.html.

[3] *Billsus D., Pazzani M. J.* Learning collaborative information filters // Proc. 15th International Conf. on Machine Learning. — Morgan Kaufmann, San Francisco, CA, 1998. — Pp. 46–54. citeseer.ist.psu.edu/billsus98learning.html.

[4] *Hofmann T.* Latent semantic models for collaborative filtering // *ACM Transactions on Information Systems.* — 2004. — Vol. 22, no. 1. — Pp. 89–115. www.cs.brown.edu/∼th.

[5] Indexing by latent semantic analysis / S. Deerwester, S. T. Dumais, G. W. Furnas et al. // *Journal of the American Society for Information Science.* — 1990. — Vol. 41. — Pp. 391–407.

[6] *Srebro N., Rennie J. D. M., Jaakkola T. S.* Maximum-margin matrix factorization // Advances in Neural Information Processing Systems 17. — MIT Press, 2005. — Pp. 1329–1336.

[7] *Hofmann T., Puzicha J.* Latent class models for collaborative filtering // International Joint Conference in Artificial Intelligence. — 1999. www.cs.brown.edu/∼th.

[8] *V.A. Leksin K. V.* The overfitting in probabilistic latent semantic models // Proceedings of 9th International Conference on Pattern Recognition and Image Analysis: New Information Technologies. — 2008. — Pp. 393–396.

[9] GroupLens: An open architecture for collaborative filtering of netnews / P. Resnick, N. Iacovou, M. Suchak et al. // Proceedings of ACM 1994 Conference on Computer Supported Cooperative Work. — Chapel Hill, North Carolina: ACM, 1994. — Pp. 175–186. citeseer.ist.psu.edu/resnick94grouplens.html.

[10] *Marlin B.* Collaborative filtering: A machine learning perspective: Ph.D. thesis / Master's thesis, University of Toronto. — 2004. citeseer.ist.psu.edu/marlin04collaborative.html.

[11] *G. Adomavicius A. T.* Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions // *IEEE Transactions on Knowledge and Data Engineering.* — 2005. — Vol. 17, no. 6.

[12] *Grcar M.* User profiling: Collaborative filtering // SIKDD 2004 at multiconference IS 12-15 Oct 2004, Ljubljana, Slovenia. — 2004.

[13] Выявление и визуализация метрических структур на множествах пользователей и ресурсов Интернет / К. В. Воронцов, К. В. Рудаков, В. А. Лексин, А. Н. Ефимов // *Искусственный Интеллект.* — 2006. — С. 285–288. www.ccas.ru/frc/papers/voron05yandex.pdf. .

[14] Generative models for cold-start recommendations / A. I. Schein, A. Popescul, L. H. Ungar, D. M. Pennock // the SIGIR'01 Workshop on Recommender Systems. — 2001. citeseer.ist.psu.edu/schein01generative.html.

[15] *Dellaert F.* The expectation maximization algorithm: Tech. rep.: 2002.

[16] *Gaussier E., Goutte C.* Relation between plsa and nmf and implications // SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval. — New York, NY, USA: ACM, 2005. — Pp. 601–602.

[17] *Воронцов К. В.* Предварительная обработка данных для решения специального класса задач распознавания // *ЖВМ и МФ.* — 1995. — Т. 35, № 10. — С. 1565–1575. www.ccas.ru/frc/papers/voron95jvm.pdf. .