УДК 519.226.3
Лексин В.А.[1]
[1] Московский физико-технический институт

Критерии ветвления в иерархическом вероятностном латентном семантическом анализе

Рассмотрим задачу, в которой на основе имеющихся данных о поведении клиентов в виде последовательности записей «клиент u выбрал ресурс r» требуется выявить потребности и интересы клиентов, а также построить адекватные функции сходства, позволяющие находить схожих клиентов и схожие ресурсы. Эти задачи решаются, в частности, методами вероятностного латентного семантического анализа (probabilistic latent semantic analysis, PLSA) [1,2], который выявляет скрытые тематические профили, характеризующие каждого клиента и каждый ресурс. Большинство из известных методов PLSA строят «плоские» профили в виде числовых векторов, не учитывающих объективно существующие иерархические взаимосвязи между тематиками. В данной работе рассматривается один из иерархических методов, и для него предлагается новый критерий расщепления компонент профиля.

Пусть заданы множество клиентов U, множество ресурсов R, и имеются данные о посещениях в виде множества пар $D = (u_i, r_i)_{i=1}^N \subset U \times R$. 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 наблюдаемому протоколу D найти скрытые профили клиентов $\{p_{tu}, t \in T\}, u \in U$ и ресурсов $\{q_{tr}, t \in T\}, r \in R$.

Пусть множество тем T имеет иерархическую структуру, то есть представляет собой дерево, в узлах которого располагаются темы. Пусть на слое $m-1$ дерева располагается тема t_{m-1}, которая содержит подтемы t_{mi} на слое m. Тогда будем полагать, что тема t_{m-1} представлена совокупностью подтем t_{mi} с весами $p(t_{mi} \mid t_{m-1})$. Причем параметр $p(t_{mi} \mid t_{m-1}) \equiv 0$, если тема t_{mi} не является подтемой для t_{m-1} (рис. 1). Безусловная вероятность темы $\hat{t}_l = (t_1, t_2, ..., t_l)$, где

$t_{m-1} = Pa(t_m)$ (родитель t_m), $m = 2,...,l$, l — номер слоя, будет равна $p(\hat{t}_l) \equiv \prod_{m=1}^{l} p(t_m \mid t_{m-1})$, $p(t_1 \mid t_0) \equiv p(t_1)$.

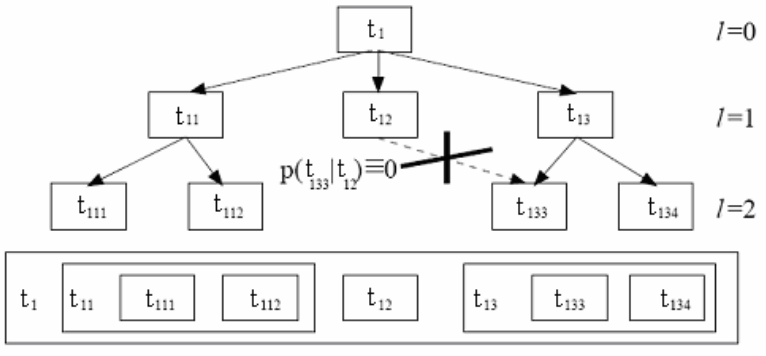

Рис.1. Пример иерархического профиля.

Обозначим через $H(t_m \mid t_{m-1}, u, r)$ вероятность того, что пользователь u и ресурс r принадлежат дочерней теме t_m, при условии того, что они принадлежат ее родительской теме t_{m-1}. Задача заключается в том, чтобы оценить эти скрытые вероятности по исходным данным. Если скрытые вероятности независимы, то $H(\hat{t}_l \mid u, r) \equiv \prod_{m=1}^{l} H(t_m \mid t_{m-1}, u, r)$, где $H(t_1 \mid t_0, u, r) \equiv p(t_1 \mid u, r)$. Тогда вероятность выбора ресурса r пользователем u:

$$p(u, r) = \sum_{t_l} p(t_l) p(u, r \mid t_l).$$

Это модель иерархического PLSA [3]. Если количество слоев в дереве равно одному, то получается модель сводится к обычному «плоскому» PLSA [1,2].

В этих моделях для оценки профилей используется принцип максимума правдоподобия (МП) $\sum_{i=1}^{N} \ln p(u_i, r_i) \to \max$ и применяется EM-алгоритм, в котором итерационно повторяются два шага. На E-шаге оцениваются скрытые переменные $H(t \mid u, r)$ — апостериорные вероятности тем t при реализации посещения (u, r). На M-шаге вычисляются наиболее правдоподобные профили, причём, благодаря введению скрытых переменных, их удаётся вычислять эффективно по аналитически выведенным формулам.

Для автоматического формирования структуры тематического дерева необходим критерий ветвления тем в профиле. Идея заключается в том, чтобы оценить эффективную длину выборки $L(t_l)$, по которой сформирована каждая тема в профиле. Затем, задав пороги L_1 и L_2, легко получить критерий расщепления или слияния темы: если $L(t_l) > L_2$, то тему можно расщеплять на подтемы; если $L(t_l) < L_1$, то тему можно сливать с родительской.

С каждой записью протокола связан набор скрытых переменных $H(t_l \mid t_{l-1}, u, r)$, причём $\sum_l H(t_l \mid t_{l-1}, u, r) = 1$, а $\sum_{ur \in D} \sum_l H(t_l \mid t_{l-1}, u, r) = N$ — длина выборки. Если взять $L(t_l) = \sum_{ur} H(t_l \mid t_{l-1}, u, r)$, то это будет «виртуальная» длина протокола, относящегося к теме t_l.Предлагаемый критерий ветвления основан на естественном требовании, чтобы темы были представлены в профилях «равномерно», т.е. оценивались по подвыборкам примерно одинаковой длины. Если на какую-то тему приходится слишком большая длина протокола, значит, об этой теме известно уже много, и можно выделять в ней подтемы.

Описанный метод оценки эффективной длины выборки тестировался на данных поисковой машины Яндекс и на данных тематического форума в качестве метода оптимизации количества тем в плоском симметризованном PLSA [1].

В случае анализа сходства текстовых сообщений, в качестве субъектов (пользователей) выступают тексты или сообщения, а в качестве объектов (ресурсов) — слова, которые в этих сообщениях содержатся. Для эксперимента были выбраны 3833 сообщений в тематическом форуме и 1472 ключевых слова. Строились профили различной длины, и была произведена оптимизация количества тем по среднеквадратичному отклонению эффективной длины выборки $L(t_l)$ от среднего значения. Оптимальным оказалось количество тем 55 с разбросом эффективной длины выборки 11%. На данных поисковой машины, которые представляли собой протокол переходов 7292 пользователей на документы (ресурсы), выданные в результатах поиска (всего были выбраны 1024 наиболее посещаемых ресурсов), оптимальным оказалось количество тем 15 с разбросом эффективной длины выборки 12%.

Работа выполнена при поддержке РФФИ, проект 08-07-00422.

Литература

[1] *Hofmann T.* Latent Semantic Models for Collaborative Filtering // ACM Transactions on Information Systems. — 2004. — V. 22, N. 1 — P. 89–115.

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

[3] *Vinokourov A., Girolami M.* A Probabilistic Framework for the Hierarchic Organisation and Classification of Document Collections // Information Processing and Management. — 2002.