Критерии ветвления в иерархическом вероятностном латентном семантическом анализе

Постановка задачи

Пусть заданы:

U – множество клиентов,

R – множество ресурсов,

$D = (u_i, r_i)_{i=1}^{N} \subset U \times R$ – протокол пользования.

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

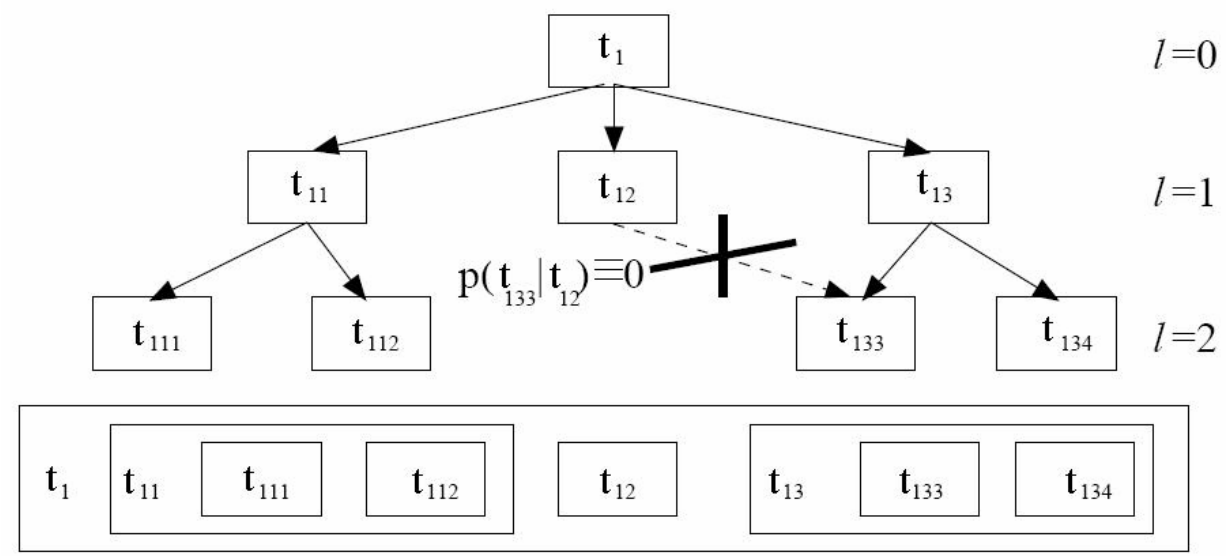
Постановка задачи

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

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

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

Иерархический профиль



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}.

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 $p(\hat{t}_l) \equiv \prod_{m=1}^{l} p(t_m \mid t_{m-1})$, $p(t_1 \mid t_0) \equiv p(t_1)$.

Скрытые вероятности

Обозначим:

$H(t_m \mid t_{m-1}, u, r)$ – 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 $H(t_1 \mid t_0, u, r) \equiv H(t_1 \mid u, r)$.

Иерархический вероятностный латентный семантический анализ, HPLSA

Вероятность выбора ресурса r пользователем u:

$$p(u,r) = \sum_{\hat{t}_l} p(\hat{t}_l)\, p(u,r \mid \hat{t}_l)\,.$$

По формуле Байеса $p(\hat{t}_l)\, p(u \mid \hat{t}_l) = p(\hat{t}_l \mid u)\, p(u)$. Тогда, опираясь на гипотезу независимости выбора пользователями ресурсов, можем записать:

$$p(u,r \mid \hat{t}_l) = p(u \mid \hat{t}_l) q(r \mid \hat{t}_l)$$

Для нахождения профилей применим принцип максимума правдоподобия (МП):

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

где максимум берется по параметрам $p(u \mid t_l)$, $q(r \mid t_l)$ и $p(t_l \mid t_{l-1})$.

EM-алгоритм

E-шаг: $H(t_l \mid t_{l-1}, u, r) = \dfrac{p(t_l \mid t_{l-1})\, p(u, r \mid \hat{t}_l)}{\sum_{t'_l} p(t'_l \mid t'_{l-1})\, p(u, r \mid \hat{t}'_l)}$.

M-шаг: $p(u \mid \hat{t}_l) = \dfrac{\sum_r f_{ur} H(\hat{t}_l \mid u, r)}{\sum_u \sum_r f_{ur} H(\hat{t}_l \mid u, r)}$, $q(r \mid \hat{t}_l) = \dfrac{\sum_u f_{ur} H(\hat{t}_l \mid u, r)}{\sum_u \sum_r f_{ur} H(\hat{t}_l \mid u, r)}$,

$$p(t_l \mid t_{l-1}) = \frac{\sum_u \sum_r f_{ur} H(t_l \mid t_{l-1}, u, r)}{\sum_u \sum_r f_{ur}}.$$

При этом профили пользователей и ресурсов выражаются по формулам:

$$p(\hat{t}_l \mid u) = \frac{p(\hat{t}_l)\, p(u \mid \hat{t}_l)}{\sum_{t'_l} p(\hat{t}'_l)\, p(u \mid \hat{t}'_l)}, \quad q(\hat{t}_l \mid r) = \frac{p(\hat{t}_l)\, q(r \mid \hat{t}_l)}{\sum_{t'_l} p(\hat{t}'_l)\, q(r \mid \hat{t}'_l)}.$$

Критерий ветвления тем в профиле

$\sum_{u,r\in D}\sum_{t_l} H(\hat{t}_l \mid u,r) = N$ — длина выборки.

$L(\hat{t}_l) = \sum_{ur} H(\hat{t}_l \mid u,r)$, – «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.

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.

Эксперименты

Анализ сходства текстовых сообщений:

- субъекты (пользователи) – тексты или сообщения,
- объекты (ресурсы) — слова, которые в этих сообщениях содержатся.

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%.

Эксперименты

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%.

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.