

Итеративное улучшение тематической модели с обратной связью от пользователя

Алексей Ильич Горбулев

Московский физико-технический институт

Курс: Моя первая научная статья/Группа Б05-021а

Эксперт: д. ф.-м. н. К. В. Воронцов

Консультант: В. А. Алексеев

2023

Цель исследования

Мотивация: тематические модели неустойчивы, неполны

Цель исследования: получить *интерпретируемую* тематическую модель за некоторое число итераций

Метод: итеративное улучшение тематической модели с использованием регуляризаторов пользовательской разметки тем на релевантные, нерелевантные и «мусорные»

- ▶ Alekseev V. et al. "TopicBank: Collection of coherent topics using multiple model training with their further use for topic model validation"
- ▶ Victor Bulatov, Evgeny Egorov, Eugenia Veselova, Darya Polyudova, Vasiliy Alekseev, Alexey Goncharov, Konstantin Vorontsov. "TopicNet: Making Additive Regularisation for Topic Modelling Accessible"
- ▶ Воронцов К.В. "Вероятностное тематическое моделирование: теория, модели, алгоритмы и проект BigARTM"

О тематическом моделировании и ARTM

► D — множество (коллекция) документов

► W — множество термов

Термами могут быть слова в нормальной форме, словосочетания.

Каждый документ $d \in D$ представляет собой последовательность термов.

► T — множество тем

Как правило, количество тем $|T|$ заранее задано.

При построении вероятностной тематической модели

$$p(w \mid d) = \sum_{t \in T} p(w \mid t) p(t \mid d) = \sum_{t \in T} \varphi_{wt} \theta_{td}$$

в подходе ARTM происходит максимизация \log правдоподобия с k регуляризаторами R_i :

$$\sum_{d,w} n_{dw} \ln \sum_t \varphi_{wt} \theta_{td} + \sum_{i=1}^k \tau_i R_i(\Phi, \Theta) \rightarrow \max_{\Phi, \Theta}$$

О тематическом моделировании и ARTM

Далее применяется EM-алгоритм.

► E-шаг:

$$p_{tdw} = \text{norm}_{t \in T}(\varphi_{wt} \theta_{td})$$

► M-шаг:

$$\varphi_{wt} = \text{norm}_{w \in W}(n_{wt} + \varphi_{wt} \frac{\partial R}{\partial \varphi_{wt}})$$

$$\theta_{td} = \text{norm}_{t \in T}(n_{td} + \theta_{td} \frac{\partial R}{\partial \theta_{td}})$$

Постановка задачи

Пусть D — коллекция документов, количество тем $|T|$ задано заранее.

После обучения базовой тематической модели M_0 каждая из тем $t \in T$ отнесена пользователем в одну из трёх категорий:

- ▶ T_+ (релевантные, *имеющие отношение к исследованию.*)
- ▶ T_0 (нерелевантные, *дублирующие релевантные*)
- ▶ T_- («мусорные», *не имеющие отношение к исследованию*)

После обучения новой тематической модели M_1 $|T_+|$ должно увеличиться, и должно быть сохранено как можно больше тем из T_+ , а $|T_-|$ должно уменьшиться. Процесс продолжается итеративно.

На каждой итерации:

- ▶ С помощью регуляризатора сглаживания

$$R(\Phi, \Theta) = \beta_0 \sum_{t \in T_0} \sum_{w \in W} \beta_{wt} \ln \varphi_{wt} + \alpha_0 \sum_{d \in D_0} \sum_{t \in T} \alpha_{td} \ln \theta_{td}$$

зафиксировать столбцы матрицы Φ , соответствующие релевантным темам, используя с достаточно большим коэффициентом

- ▶ Для выявления новых релевантных тем использовать регуляризатор декоррелирования, используя матрицу $\tilde{\Phi}$ предыдущей модели:

$$R(\Phi) = -\tau \sum_{t \in T_+} \sum_{s \in T_-} \sum_{w \in W} \varphi_{wt} \tilde{\varphi}_{ws} \rightarrow \max$$

$$\varphi_{wt} = \text{norm}_{w \in W} \left(n_{wt} - \tau \varphi_{wt} [t \in T_+] \sum_{s \in T_-} \tilde{\varphi}_{ws} \right)$$

Вычислительный эксперимент

В качестве коллекции текстов используется набор из 16 449 новостей, опубликованных на сайте Lenta.ru в период с мая по август 2008-го года. Предполагается разделение на 100 тем.

Предобработка:

- ▶ Заголовок и текст каждой новости разбиваются на токены, далее происходит лемматизация.
- ▶ Далее по PMI отбирается 10 000 биграмм, которые характеризуют коллекцию текстов.

Базовая модель: TopicNet, 50 предметных тем, без регуляризаторов на темы

Новая модель: TopicNet, 50 предметных тем, регуляризатор сглаживания для тем T_+ с $\tau = 10^9$, регуляризатор декоррелирования для тем из T_-

Каждая последующая модель строится аналогично, используя данные модели на предыдущей итерации.

Вычислительный эксперимент

Внешний критерий: количество тем в T_+ , T_0 , T_- .

Чем больше $|T_+|$ и меньше $|T_-|$, тем лучше.

Внутренние критерии:

▶ Перплексия

$$\mathcal{P}_m(D; p) = \exp \left(-\frac{1}{n_m} \sum_{d \in D} \sum_{w \in W^m} n_{dw} \ln p(w \mid d) \right)$$

▶ Разреженность матрицы Φ

▶ Средняя контрастность тем, где контрастность темы определяется как

$$\text{con}_t = \frac{1}{|W_t|} \sum_{w \in W_t} p(t \mid w), \quad W_t = \{w \in W \mid \varphi_{wt} > \frac{1}{|W|}\}$$

Вычислительный эксперимент

Модель	$ T_+ $	$ T_0 $	$ T_- $
M_0	5	1	44
M_1	7	1	42
M_2	8	1	41
M_3	8	1	41

Таблица: Данные по группам по пользовательской разметке

На каждой итерации удалось сохранить ранее найденные релевантные темы.

Вычислительный эксперимент

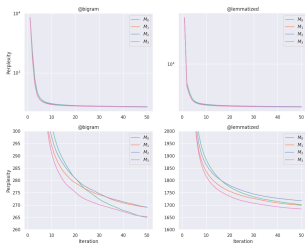


Рис.: Перплексия

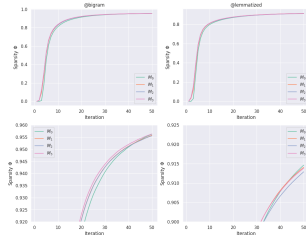


Рис.: Разреженность Φ

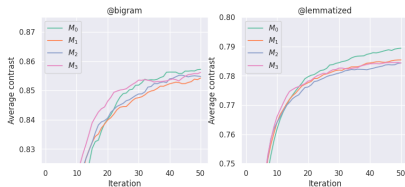


Рис.: Средняя контрастность тем

Результаты:

- ▶ предложен метод итеративного улучшения тематической модели,
- ▶ показано, как использовать регулятор сглаживания для сохранения тем из T_+ ,
- ▶ предложен регуляризатор декоррелирования тем из T_- .