УДК: 004.852

Итеративное улучшение тематической модели с обратной связью от пользователя

А. И. Горбулев, В. А. Алексеев, К. В. Воронцов
Московский физико-технический институт
(национальный исследовательский университет)

Одним из методов анализа текстов, который применяется в том числе в социологических исследованиях [2] и активно развивается в последнее время [1], является тематическое моделирование. В то же время, не все темы могут оказаться *релевантными* в контексте проводимого исследования. Часть документов, которая по содержанию релевантны, могут быть отнесены как к *нерелевантной* теме, которая дублирует по содержанию релевантную, так и к *«мусорной»* теме, которая не имеет отношения к исследованию, что негативно влияет на качество исследования.

Целью данного исследования является построение тематической модели как результата итеративного улучшения в процессе обучения в процессе обучения нескольких моделей.

Пользователь относит каждую из тем к одной из трёх категорий: релевантные, нерелевантные и «мусорные» темы. Улучшение модели, которое основывается на пользовательской разметке, способствует сохранению ранее найденных релевантных тем и выделению пользователем новых релевантных тем, а также уменьшению числа «мусорных тем».

Для решения задачи используются в том числе и методы аддитивной регуляризации тематических моделей (ARTM), реализованные в библиотеках с открытым кодом `BigARTM` и `TopicNet`. В качестве набора текстовых данных используется коллекция, основанная на новостях, опубликованных на сайте Lenta.ru в период с мая по август 2008 года.

Пусть D — коллекция текстов, W — множество термов. Среди термов могут быть как ключевые слова, так и словосочетания [3]. Каждый документ $d \in D$ представим в виде последовательности n_d термов $(w_1, \ldots, w_{n_d})$ из множества W [5]. Предполагается конечное множество тем T. Коллекция документов D рассматривается выборка из дискретного распределения $p(w, d, t)$ на конечном множестве $W \times D \times T$ [3]. Согласно формуле полной вероятности и гипотезе условной независимости, распределение термов в документе $p(w \mid d)$ описывается вероятностной смесью распределений термов в темах $\varphi_{wt} = p(w \mid t)$ с весами $\theta_{td} = p(t \mid d)$ следующим образом: [4]

$$p(w \mid d) = \sum_{t \in T} p(w \mid t, d) p(t \mid d) = \sum_{t \in T} p(w \mid t) p(t \mid d) = \sum_{t \in T} \varphi_{wt} \theta_{td} \tag{1}$$

Задача тематического моделирования состоит в нахождении по коллекции документов D параметров φ_{wt} и θ_{td}, приближающих частотные оценки условных вероятностей $\widehat{p}(w \mid d)$. Так как $|T|$ обычно намного меньше, чем $|W|$ и $|D|$, то находится низкоранговое стохастическое матричное разложение [3]

$$F \approx \Phi\Theta \tag{2}$$

где $F = (\widehat{p}_{wd})_{|W| \times |D|}$ — матрица частот терм в документах, $\Phi = (\varphi_{wt})_{|W| \times |T|}$ — матрица термов тем, $\Theta = (\theta_{td})_{|T| \times |D|}$ — матрица тем документов.

Предполагается T^i — множество тем на итерации $i \in \mathbb{N}$, $T^i_+ \subset T^i$ — подмножество релевантных тем с точки зрения пользователя, $T^i_0 \subset T^i$ — подмножество нерелевантных тем с точки зрения пользователя, $T^i_- \subset T^i$ — подмножество «мусорных» тем с точки

зрения пользователя, при этом $T^i = T^i_+ \sqcup T^i_0 \sqcup T^i_-$, M_i — состояние модели на итерации i. Тогда итеративное улучшение модели M_i состоит в построении модели M_{i+1}, такой, чтобы множество тем T^{i+1} удовлетворяло следующим требованиям:

$$T^i_+ \subset T^{i+1}_+, \ \left|T^{i+1}_-\right| \leq \left|T^{i+1}_-\right|$$

Предполагается при обучении новой модели M_{i+1} использовать аддитивную регуляризацию тематических моделей (ARTM) и выполнить следующее:

1. Использовать альтернативное значение параметра, отвечающего за генерацию случайного начального приближения;

2. С помощью регуляризатора сглаживания

$$R(\Phi, \Theta) = \beta_0 \sum_{t \in T_+} \sum_{w \in W} \widetilde{\varphi}_{wt} \ln \varphi_{wt} \tag{3}$$

   и матрицы $\widetilde{\Phi}$ модели M_i зафиксировать столбцы матрицы Φ, соответствующие релевантным темам, используя с достаточно большим коэффициентом, $\beta_0 \gg 1$;

3. Чтобы способствовать выявлению новых релевантных тем, предлагается использовать регуляризатор декоррелирования, используя матрицу $\widetilde{\Phi}$ из модели M_i:

$$R(\Phi) = -\tau \sum_{t \in T_- \cup T_0} \sum_{s \in T_-} \sum_{w \in W} \varphi_{wt} \widetilde{\varphi}_{ws} \to \max \tag{4}$$

$$\varphi_{wt} = norm_{w \in W} \left(n_{wt} - \tau \varphi_{wt} [t \in T_- \cup T_0] \sum_{s \in T_-} \widetilde{\varphi}_{ws} \right) \tag{5}$$

Проведённые вычислительные эксперименты (https://github.com/intsystems/2023-Project-131/tree/master/code) показывают увеличение числа релевантных тем и сохранение ранее найденных тем с помощью регуляризатора (3).

Список литературы

[1] Jedidiah Aqui and Michael Hosein. Mobile ad-hoc networks topic modelling and dataset querying. In *2022 IEEE 2nd International Conference on Mobile Networks and Wireless Communications (ICMNWC)*, pages 1–6, 2022.

[2] Paul DiMaggio, Manish Nag, and David Blei. Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of u.s. government arts funding. *Poetics*, 41(6):570–606, 2013. Topic Models and the Cultural Sciences.

[3] Konstantin Vorontsov and Anna Potapenko. Additive regularization of topic models. *Machine Learning*, 101:1–21, 12 2014.

[4] К. В. Воронцов. Вероятностное тематическое моделирование: теория, модели, алгоритмы и проект bigartm.

[5] К. В. Воронцов. Аддитивная регуляризация тематических моделей коллекций текстовых документов. volume 456, pages 268–271, 2014.