Итеративное улучшение тематической модели с обратной связью от пользователя

Горбулев Алексей Ильич
gorbulev.ai@phystech.edu

Алексеев Василий Антонович
vasiliy.alekseyev@phystech.edu

Воронцов Константин Вячеславович
vokov@forecsys.ru

Аннотация

В работе представлен метод тематического моделирования с использованием обратной связи от пользователя. Обратная связь заключается в определении принадлежности темы, полученной при тематическом моделировании, к одной из трёх категорий: релевантная, нерелевантная, «мусорная». Основная задача состоит в улучшении базовой модели, которое заключается в выделении новых релевантных тем при сохранении выделенных тем и уменьшении числа «мусорных» тем. В работе предлагается решение с использованием библиотек тематического моделирования и регуляризаторов сглаживания и декоррелирования. Вычислительный эксперимент проводится на текстовой коллекции, основанной на новостях сайта Lenta.ru.

Ключевые слова: Тематическое моделирование · ARTM · Обработка естественного языка

1 Введение

Одним из методов анализа текстов, который применяется в том числе в социологических исследованиях [3] и активно развивается в последнее время [1], является тематическое моделирование. Тематическая модель помогает оценить вероятность принадлежности текста к каждой из полученных тем. Именно вероятностное тематическое моделирование использовалось при исследовании распространения информации о пандемии COVID-19 в Хорватии [2], освещения в средствах массовых информации Литвы климатических изменений [4].

Однако задача построения вероятностной тематической модели имеет бесконечно много решений [6] вследствие некорректной постановки. С целью ограничения количества решений вводятся регуляризаторы. Например, декорреляция способствует улучшению когерентности тем [5], разреживание способствует обнулению части элементов матриц.

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 обновляемой тематической модели с использованием обратной связи от пользователя, найденные релевантные темы которой охватывают в совокупности большую часть коллекции документов. 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, которые включают в себя регуляризаторы декоррелирования и сглаживания. Именно подход ARTM способствует оптимизации моделей по сумме нескольких критериев [7], что помогает учитывать особенности коллекции текстов и ограничить количество решений задачи тематического моделлирования. В качестве набора

текстовых данных используется коллекция, основанная на новостях, опубликованных на сайте Lenta.ru в период с 1999 по 2019 годы.

2 Постановка задачи

Пусть D — коллекция текстов, W — множество термов. Среди термов могут быть как ключевые слова, так и словосочетания [5]. Каждый документ $d \in D$ представим в виде последовательности n_d термов $(w_1, \ldots, w_{n_d})$ из множества W [7]. Предполагается конечное множество тем T. Коллекция документов D рассматривается выборка из дискретного распределения $p(w, d, t)$ на конечном множестве $W \times D \times T$ [5]. 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: [6]

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

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 [5]

$$F \approx \Phi\Theta$$

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 $i \in \mathbb{N}$, $T_i^1 \subset T_i$ — подмножество релевантных тем с точки зрения пользователя, $T_i^2 \subset T_i$ — подмножество нерелевантных тем с точки зрения пользователя, $T_i^3 \subset T_i$ — подмножество «мусорных» тем с точки зрения пользователя, при этом $T_i = \sqcup_{k=1}^{3} T_i^k$, 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:

$$T_i^1 \subset T_{i+1}^1, \ \left|T_{i+1}^3\right| \leq \left|T_i^3\right|$$

3 Метод

Предполагается при обучении новой модели M_{i+1} выполнить следующее:

1. При инициализации матрицы Φ зафиксировать столбцы, соответствующие релевантным темам T_i^1;
2. При обучении использовать аддитивную регуляризацию тематических моделей (ARTM):
   (a) Для уменьшения числа дублирующих друг друга тем использовать регуляризатор декоррелирования, действующий на незафиксированные столбцы матрицы Φ, матрицу из которых обозначим через Φ_{new}:
   $$R(\Phi_{new}) = -\frac{\tau}{2} \sum_{t \in (T_i^2 \cup T_i^3)} \sum_{s \in T_i^3 \setminus t} \sum_{w \in W} \varphi_{wt} \varphi_{ws}$$
   (b) Для улучшения интерпретируемости тем использовать регуляризаторы сглаживания и разреживания относительно Φ_{new}.

4 Вычислительный эксперимент

Целью данного вычислительного эксперимента является нахождение способа улучшения тематической модели. После обучения базовой модели полученные темы распределяются пользователем на три категории: релевантные, нерелевантные, "мусорные". Далее используется несколько метрик, в том числе разреженность. Новая модель строится добавлением новых регуляризаторов и изменением параметров. Сравнивается количество тем в каждой категории и новые значения метрик.

4.1 Описание данных

Описан базовый вариант.

Для обучения тематических моделей использовались новости, опубликованные на сайте Lenta.ru с 1 января по 31 марта 2002 года. Каждая из 5492 новостей распределена по одной из 8 макротем: «Из жизни», «Интернет и СМИ», «Культура», «Мир», «Россия», «Силовые структуры», «Спорт», «Экономика». Далее происходит предобработка данных с помощью токенизации, лемматизации, выделения биграмм, приведения текста в формат, совместимым с библиотекой Vowpal Wabbit. Каждая тема характеризуется пользователем по 5 ключевым словам и 5 ключевым биграммам.

4.2 Базовая модель

В качестве базовой модели M_0 используется тематическая модель с регуляризатором декоррелирования, применяемым относительно слов, прошедших лемматизацию Всего используется 50 специальных тем и 1 фоновая. Далее используется тематическая модель M_1, использующая регуляризаторы декоррелирования и сглаживания как для лемматизированных слов, так и для биграмм.

4.3 Предварительные результаты

Среди тем, предложенных базовой моделью, были выделены 27 релевантных, 11 нерелевантных и 12 «мусорных» тем. Добавление новых регуляризаторов декоррелирования и сглаживания как для лемматизированных слов, так и для биграмм, позволило выделить 2 новые релевантные темы среди «мусорных», а также сохранить все ранее выделенные релевантные темы.

Модель	$\|T_1\|$	$\|T_2\|$	$\|T_3\|$
M_0	27	11	12
M_1	29	11	10

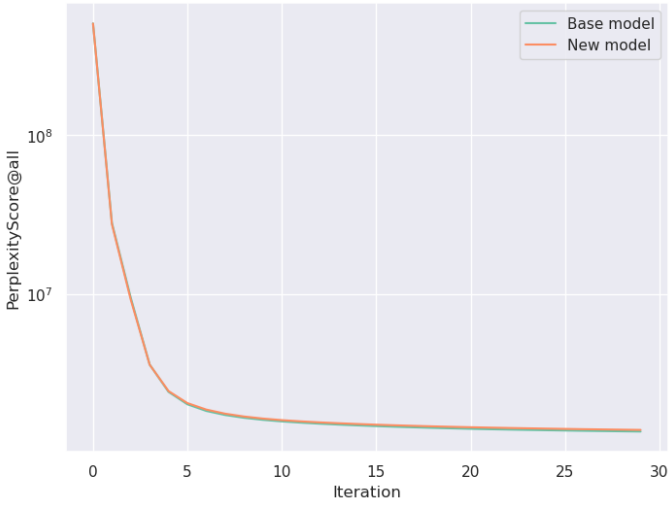

Рис. 1: Перплексия

Для новой модели разреженность выше, чем для базовой модели, что позволяет более точно определять темы с меньшим количеством возможных конфликтов. Изменились и размеры ядер тем, что отразилось и на теме 41, признанной новой релевантной. При этом происходит крайне незначительная потеря перплексии, что является обычной ситуацией [5].

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.

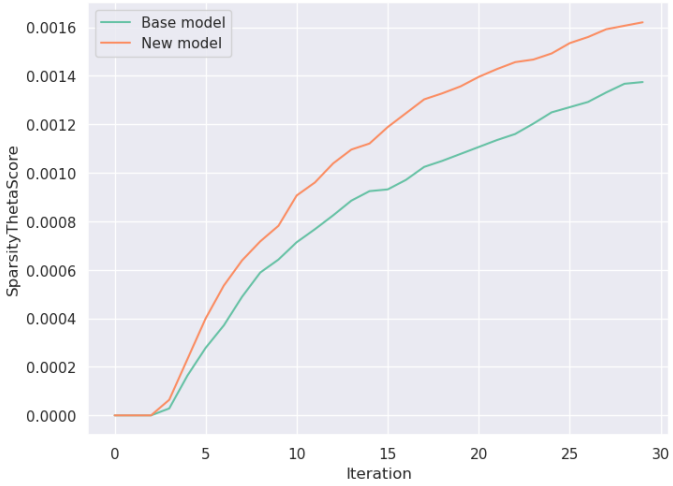

Рис. 2: Разреженность матрицы Θ

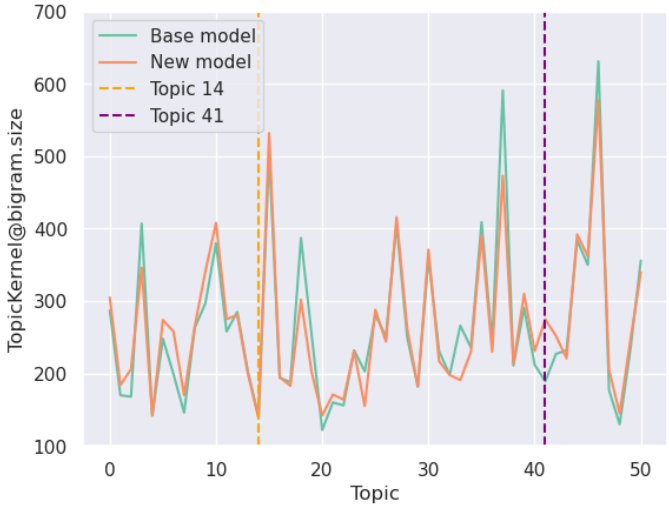

Рис. 3: Итоговый размер ядер тем

[2] Petar Kristijan Bogović, Ana Meštrović, Slobodan Beliga, and Sanda Martinčić-Ipšić. Topic modelling of croatian news during covid-19 pandemic. In 2021 44th International Convention on Information, Communication and Electronic Technology (MIPRO), pages 1044–1051, 2021.

[3] 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.

[4] Florian Rabitz, Audronė Telešienė, and Eimantė Zolubienė. Topic modelling the news media representation of climate change. Environmental Sociology, 7(3):214–224, 2021.

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

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

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