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 [4] 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 [6].

Однако задача построения вероятностной тематической модели имеет бесконечно много решений [8] вследствие некорректной постановки. С целью ограничения количества решений вводятся регуляризаторы. Например, декорреляция способствует улучшению когерентности тем [7], 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 [3], которые включают в себя регуляризаторы декоррелирования и сглаживания. Именно подход ARTM способствует оптимизации моделей по сумме нескольких критериев [9], 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 мая по август 2008 года.

2 Постановка задачи

Пусть D — коллекция текстов, W — множество термов. Среди термов могут быть как ключевые слова, так и словосочетания [7]. Каждый документ $d \in D$ представим в виде последовательности n_d термов $(w_1, \ldots, w_{n_d})$ из множества W [9]. Предполагается конечное множество тем T. Коллекция документов D рассматривается выборка из дискретного распределения $p(w, d, t)$ на конечном множестве $W \times D \times T$ [7]. 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: [8]

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

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 [7]

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

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:

$$T^1_i \subset T^1_{i+1}, \ \left|T^3_{i+1}\right| \leq \left|T^3_i\right|$$

3 Метод

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. При инициализации матрицы Φ зафиксировать столбцы, соответствующие релевантным темам T^i_+, используя регуляризатор сглаживания, общий вид формулы которого [10]

$$R(\Phi, \Theta) = \beta_0 \sum_{t \in T} \sum_{w \in W} \beta_{wt} \ln \varphi_{wt} + \alpha_0 \sum_{d \in D} \sum_{t \in T} \alpha_{td} \ln \theta_{td} \to \max_{\Phi, \Theta} \tag{3}$$

   Коэффициент τ, соответствующий данному регуляризатору, должен быть достаточно большим для фиксации. В вычислительном эксперименте рассматривается значение τ, равное 10^9. При использовании другого значения параметра, ответственного за генерацию случайного начального приближения, нерелевантные темы могут не сохраниться, и может увеличиться число «мусорных» тем. Вследствие этого предлагается также зафиксировать столбцы, соответствующие нерелевантным темам T^i_0, так как тема может не полностью пересекаться с соответствующей ей релевантной.
3. Чтобы способствовать выявлению новых релевантных тем, предлагается использовать регуляризатор декоррелирования, используя матрицу $\widetilde{\Phi}$ из модели M_i:

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

$$R(\Phi) = -\tau \sum_{t \in T_+ s \in T_-} \langle \varphi_t, \widetilde{\varphi}_s \rangle \to \max \tag{5}$$

$$R(\Phi) = -\tau \sum_{t \in T_+} \left\langle \varphi_t, \sum_{s \in T_-} \widetilde{\varphi}_s \right\rangle \to \max \tag{6}$$

$$\frac{\partial R}{\partial \varphi_{wt}} = -\tau [t \in T_+] \sum_{s \in T_-} \widetilde{\varphi}_{ws} \tag{7}$$

$$\varphi_{wt} = norm_{w \in W} \left(n_{wt} - \tau \varphi_{wt} [t \in T_+] \sum_{s \in T_-} \widetilde{\varphi}_{ws} \right) \tag{8}$$

В качестве внешнего критерия предлагается использовать количество тем в T_+, T_0 и T_-. Чем больше тем в T_+ и меньше тем в T_i, тем лучше.

В качестве внутреннего критерия предлагается использовать следующие метрики:

1. Перплексия [8]
$$\mathcal{P}_m(D; p) = \exp \left(-\frac{1}{n_m} \sum_{d \in D} \sum_{w \in W^m} n_{dw} \ln p(w \mid d) \right) \tag{9}$$
2. Разреженность матрицы Φ
3. Средняя контрастность тем [8], где контрастность темы определяется как
$$con_t = \frac{1}{|W_t|} \sum_{w \in W_t} p(t \mid w) \tag{10}$$
$$W_t = \{ w \in W \mid \varphi_{wt} > \frac{1}{|W|} \} \tag{11}$$

4 Вычислительный эксперимент

Вычислительный эксперимент был произведён на языке Python с использованием библиотек с открытым исходным кодом BigARTM и TopicNet.

4.1 Описание данных

Для обучения тематических моделей использовалась коллекция новостей, опубликованных на сайте Lenta.ru с 1 мая по 31 августа 2008 года. Каждая из 16449 новостей распределена по одной из 11 макротем:

1. «Бывший СССР»
2. «Дом»
3. «Из жизни»
4. «Интернет и СМИ»
5. «Культура»
6. «Мир»
7. «Наука и техника»
8. «Россия»
9. «Силовые структуры»
10. «Спорт»
11. «Экономика»

Новости, относящиеся к одной из макротем, могут быть отнесены как к релевантной, так и к «мусорной» теме. В связи с этим предлагается в качестве параметра, отвечающее за количество предметных тем $|T|$, использовать значение, равное 50.

4.2 Предобработка

Заголовки и тексты каждой новости были разбиты на токены, которые представляют собой отдельные слова. Далее посредством морфологического анализатора pymorphy2 [5] была произведена лемматизация токенов, которые не вошли в список стоп-слов. Из двух подряд идущих лемматизированных токенов были составлены биграммы. Далее по метрике PMI (pointwise mutual information, поточечная взаимная информация) были отобраны 10000 биграмм, которые характеризуют каждый документ из коллекции. Далее произошло преобразование в формат, совместимый с Vowpal Wabbit, чтобы обеспечить работу с TopicNet.

Суммарно предобработка прошла за 11 минут 14 секунд.

4.3 Базовая модель

В качестве базовой модели M_0 используется тематическая модель, не имеющая регуляризаторов, действующих на столбцы матрицы Φ, соответствующих специальным темам, и имеющая регуляризаторы сглаживания, действующие на столбцы матриц Φ и Θ, соответствующие фоновой теме, имеющие коэффициенты, равные 0. Используются две модальности:

1. По лемматизированным токенам (@lemmatized), коэффициент 1.0;
2. По биграммам (@bigram), коэффициент 0.8.

В качестве значения параметра seed, отвечающего за генерацию случайного начального приближения, используется 21. Обучение модели M_0 происходит в течение 50 итераций.

4.4 Создание новой модели

Пусть известна пользовательская разметка на основе тематической модели M_{i-1}, а также параметры модели M_{i-1}. Создание модели M_i происходит с соблюдением следующих правил:

1. Значение параметра seed увеличивается на 21;
2. С помощью регуляризатора сглаживания происходит фиксация тем из T_+^{i-1} и T_0^{i-1};
3. Создаётся регуляризатор декоррелирования (4) для тем из T_+^{i-1} и T_-^{i-1};
4. Остальные параметры сохраняются.

Одним из недостатков тематического моделирования в целом является неполнота моделей и отсутствие однозначного решения задачи [8], а также зависимость от случайности. Одним из способов избавления от случайности является использование различных значений параметров seed при обучении разных версий моделей.

4.5 Результаты

После обучения базовой модели M_0 пользователем было выделено 5 релевантных тем:

1. 12 (Дмитрий Медведев и его международные контакты)
2. 21 (события на Кавказе в августе 2008 года)
3. 29 (Украина, политика и ситуация по переговорам по газовой сделке)
4. 30 (президентские выборы в США, 2008)
5. 35 (Путин в качестве премьер-министра)

Была выделена 1 нерелевантная тема 31, в большинстве своём дублирующая тему 21. Была создана модель M_1, содержащийся в ней регуляризатор фиксации тем помог сохранить ранее найденные релевантные темы. Кроме того, были выделены ещё 2 релевантные темы:

1. 14 (санкции США против Ирана)
2. 15 (выборы, ситуация в Зимбабве)

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

Таблица 1: Данные по группам по пользовательской разметке

После обучения ещё одной новой модели M_2 выяснилось, что темы из T_+^1 были сохранены, также удалось выделить релевантную тему 46 о Гаагском трибунале. После обучения M_3 разметка тем по трём категориям не изменилась по сравнению с M_2.

По обоим модальностям модели M_1 и M_2 обладают меньшей перплексией, чем M_0, что обычно происходит при использовании регуляризаторов. [7] В то же время, после 50 итераций обучения модель M_3 обладает перплексией, схожей с перплексией модели M_0.

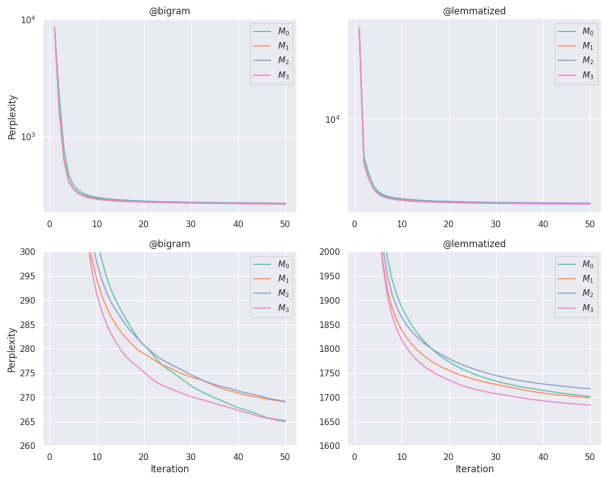

Рис. 1: Перплексия

Различия по разреженности матрицы Φ крайне незначительны от модели к модели и находятся в пределах статистической погрешности.

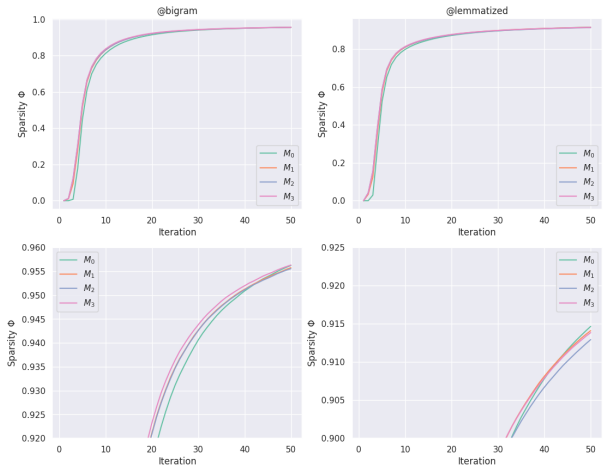

Рис. 2: Разреженность матрицы Θ

По средней контрастности тем модель M_3 показывает схожие результаты с M_0, которые являются сравнительно лучшими по модальности биграмм. С точки зрения модальности лемматизированных

слов, модель M_0 обладает более высокой средней контрастностью тем, чем остальные. В то же время, различия являются несущественными.

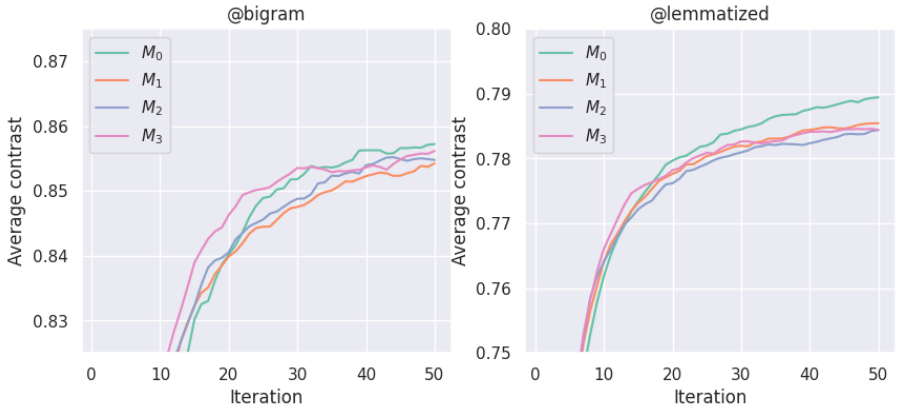

Рис. 3: Средняя контрастность тем

5 Заключение

В данной работе был предложен метод итеративного улучшения тематической модели с помощью фиксации релевантных тем через регуляризатор сглаживания и декоррелирования тем. Вычислительный эксперимент показал, что регуляризатор сглаживания, действующий на столбцы матрицы Ф, соответствующий релевантным темам, действительно способствует сохранению тем. Кроме того, удалось добиться увеличения количества релевантных тем. В дальнейшем планируется исследование не только на коллекциях новостей, но и на коллекциях специализированных текстов, чтобы проверить универсальность предложенного метода.

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] 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] Victor Bulatov, Vasiliy Alekseev, Konstantin Vorontsov, Darya Polyudova, Eugenia Veselova, Alexey Goncharov, and Evgeny Egorov. TopicNet: Making additive regularisation for topic modelling accessible. In Proceedings of the Twelfth Language Resources and Evaluation Conference, pages 6745–6752, Marseille, France, May 2020. European Language Resources Association.

[4] 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.

[5] Mikhail Korobov. Morphological analyzer and generator for russian and ukrainian languages. In Mikhail Yu. Khachay, Natalia Konstantinova, Alexander Panchenko, Dmitry I. Ignatov, and Valeri G. Labunets, editors, Analysis of Images, Social Networks and Texts, volume 542 of Communications in Computer and Information Science, pages 320–332. Springer International Publishing, 2015.

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

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

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

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

[10] К. В. Сухарева А. В., Воронцов. Построение полного набора тем вероятностных тематических моделей. In Интеллектуальные системы. Теория и приложения, volume 23, pages 7–23, 2019.