Переосмысление вероятностных тематических моделей с позиций классической не-байесовской регуляризации

Воронцов Константин Вячеславович

ФИЦ ИУ РАН и Институт ИИ МГУ, voron@mlsa-iai.ru

Аннотация

Вероятностные тематические моделирование было успешной техникой для анализа текстов почти двадцать лет, собрав в своём арсенале сотни тематических моделей и их приложений. Эта область исследований развивалась главным образом в рамках теории байесовского обучения. Долгое время возможность построения тематических моделей на более простой основе классической (не-байесовской) регуляризации оставалась недооцененной и редко используемой. Формализм Аддитивной Регуляризации Тематических Моделей (АРТМ) восполняет этот пробел. Он существенно упрощает вывод алгоритмов для обучения моделей, а также открывает новые возможности комбинирования тематических моделей путём сложения их регуляризаторов. Это делает АРТМ инструментом для синтеза композитных моделей с требуемыми свойствами и возможностями. АРТМ приводит к эффективному онлайновому алгоритму, реализованному в проекте с открытым кодом BigARTM, снабжённому расширяемой модульной библиотекой регуляризаторов для комбинирования тематических моделей. В данной статье мы идём дальше и показываем, что разнообразные тематические модели получаются как частные случаи общего итерационного процесса для максимизации гладкой функции на единичных симплексах. Полагаем, что данный подход окажется полезным не только для переосмысления вероятностного тематического моделирования, но и для построения нейросетевых тематических моделей, которые всё активнее предлагаются в последние годы.

1 Введение

Тематическое моделирование — это популярный инструмент обработки естественного языка, который активно развивается с конца 90-х годов и находит множество применений [7, 10, 28, 15]. Вероятностная тематическая модель определяет тематику коллекции текстовых документов, описывая каждую тему распределением вероятностей слов, а каждый документ — распределением вероятностей тем.

Тематическое моделирование называют мягкой кластеризацией текстов. В отличие от обычной жёсткой кластеризации, документ не относится целиком к одному кластеру, а распределяется между несколькими кластерами-темами. Тематические модели называют также моделями мягкой би-кластеризации, поскольку слова также распределяются по темам.

Задача тематического моделирования текстовой коллекции сводится к низкоранговому матричному разложению. Это некорректно поставленная задача, имеющая бесконечно много решений. Чтобы доопределить решение и сделать его устойчивым, вводятся регуляризаторы [49], накладывающие дополнительные ограничения на модель. В сложных задачах регуляризаторов может быть несколько.

Начиная с модели латентного размещения Дирихле, LDA [8], доминирующим подходом в тематическом моделировании остаётся байесовское обучение. Главный его недостаток заключается в том, что процесс байесовского вывода уникален для каждой модели, и чем сложнее модель, тем сложнее её вывод. Не существует на данный момент простых способов автоматизации этого вывода или конструирования сложных моделей из более простых. Байесовская регуляризация вводится с помощью априорных распределений, однако более удобным и общепринятым способом являются оптимизационные критерии. Многие модели предполагают априорные распределения Дирихле, что упрощает байесовский вывод благодаря свойству сопряжённости. Именно математическое удобство предопределило особую роль распределения Дирихле в тематическом моделировании, при отсутствии убедительных лингвистических обоснований. Наконец, механизмы байесовского вывода неудобно совмещать с обучением нейросетевых моделей языка [61]. Эти барьеры препятствуют широкому распространению тематического моделирования. В индустрии анализа текстов редко применяются тематические модели сложнее LDA. Сотни моделей так и остаются исследованиями для одной статьи.

Перечисленные недостатки преодолеваются в подходе аддитивной регуляризации тематических моделей (ARTM), который основан на классической не-байесовской регуляризации [52, 55]. Как показано в [31], широкий класс байесовских тематических моделей допускает переформулировку в терминах ARTM. После этого появляется возможность переносить регуляризаторы из одних моделей в другие или складывать регуляризаторы от различных моделей для получения комбинированных моделей с требуемыми свойствами. Для их обучения используется общий алгоритм, в котором регуляризаторы можно подключать как модули. Модульная технология реализована в библиотеке тематического моделирования с открытым кодом `BigARTM`, `http://bigartm.org` [53, 20]. Подчеркнём, что ARTM не является ещё одной моделью или методом — это общий подход к построению и комбинированию тематических моделей.

В данной работе предлагается ещё более общий подход. Доказана теорема о максимизации гладкой функции на единичных симплексах. Из этой теоремы выводятся итерационные EM-подобные алгоритмы для обучения тематических моделей различной структуры с про-

извольными гладкими регуляризаторами. Фактически, тематическое моделирование становится теорией единственной теоремы.

Одна итерация этого алгоритма мало отличается от градиентного шага при обучении нейронных сетей. Данное обстоятельство открывает новые возможности для построения нейросетевых тематических моделей, а также нейронных сетей с ограничениями неотрицательности и нормированности на часть векторов.

2 Теорема о максимизации гладкой функции на единичных симплексах

Введём оператор norm, преобразующий произвольный числовой вектор $(x_i)_{i\in I}$ в неотрицательный нормированный вектор:

$$p_i = \operatorname*{norm}_{i\in I}(x_i) = \frac{(x_i)_+}{\sum\limits_{k\in I}(x_k)_+}, \text{ для всех } i \in I,$$

где $(x)_+ = \max\{0, x\}$ — операция положительной срезки. Если $x_i \leqslant 0$ для всех $i \in I$, то результатом оператора norm будем считать нулевой вектор. В остальных случаях вектор $(p_i)_{i\in I}$ лежит на единичном симплексе и определяет дискретное распределение вероятностей на конечном множестве элементарных исходов I.

Теорема 2.1 *Пусть функция $f(\Omega)$ непрерывно дифференцируема по набору векторов $\Omega = (\omega_j)_{j\in J}$, $\omega_j = (\omega_{ij})_{i\in I_j}$. Если ω_j — вектор локального экстремума задачи математического программирования*

$$f(\Omega) \to \max_{\Omega}, \quad \sum_{i\in I_j} \omega_{ij} = 1, \quad \omega_{ij} \geqslant 0, \quad i \in I_j,\ j \in J$$

и если $\omega_{ij}\frac{\partial f}{\partial \omega_{ij}} > 0$ при некотором i, то ω_j удовлетворяет уравнениям

$$\omega_{ij} = \operatorname*{norm}_{i\in I_j}\left(\omega_{ij}\frac{\partial f}{\partial \omega_{ij}}\right). \tag{1}$$

Доказательство. Запишем лагранжиан оптимизационной задачи с ограничениями неотрицательности и нормированности векторов:

$$\mathscr{L}(\Omega) = f(\Omega) - \sum_{j\in J}\lambda_j\left(\sum_{i\in I_j}\omega_{ij} - 1\right) + \sum_{j\in J}\sum_{i\in I_j}\mu_{ij}\omega_{ij},$$

где множители λ_j соответствуют ограничениям нормировки, μ_{ij} — неотрицательности. Запишем условия Каруша–Куна–Таккера, приравняв нулю производные лагранжиана по параметрам модели:

$$\frac{\partial \mathscr{L}}{\partial \omega_{ij}} = \frac{\partial f}{\partial \omega_{ij}} - \lambda_i + \mu_{ij} = 0; \quad \mu_{ij}\omega_{ij} = 0. \tag{2}$$

Умножив обе части равенства (2) на ω_{ij}, получим уравнение

$$\omega_{ij}\frac{\partial f}{\partial \omega_{ij}} = \omega_{ij}\lambda_j.$$

Обозначим левую часть равенства через A_{ij}. Тогда $A_{ij} = \omega_{ij}\lambda_j$.

Согласно условию теоремы, существует i, для которого $A_{ij} > 0$. Значит, $\lambda_j > 0$. Если $\frac{\partial f}{\partial \omega_{ij}} < 0$ для некоторого i, то $\mu_{ij} = \lambda_i - \frac{\partial f}{\partial \omega_{ij}} > 0$, следовательно, $\omega_{ij} = 0$.

Объединяя уравнение $\omega_{ij}\lambda_t = A_{ij}$ при $A_{ij} > 0$ с нулевым решением $\omega_{ij} = 0$ при $A_{ij} \leqslant 0$, получим $\omega_{ij}\lambda_j = (A_{ij})_+$. Суммируя эти уравнения по i, выразим двойственную переменную: $\lambda_j = \sum_{i \in I_j} (A_{ij})_+$. Подставляя λ_j в формулу $\omega_{ij} = \frac{1}{\lambda_j}(A_{ij})_+$, получим требуемое (1).

Теорема доказана.

Для решения полученной системы уравнений удобно использовать метод простой итерации. Обновления векторов ω_j похожи на шаги градиентной максимизации $\omega_{ij} = \omega_{ij} + \eta \frac{\partial f}{\partial \omega_{ij}}$. В обоих случаях требуется вычислять градиент функции $f(\Omega)$. Отличий три: вместо аддитивного градиентного шага используется мультипликативный, после этого вектор проецируется на единичный симплекс оператором norm, и величину градиентного шага η подбирать не нужно.

Предположим, что условие применимости формулы (1) выполняется для всех векторов, и рассмотрим итерационный процесс

$$\omega_{ij}^{t+1} = \underset{i \in I_j}{\operatorname{norm}}\left(\omega_{ij}^t \frac{\partial f(\Omega^t)}{\partial \omega_{ij}^t}\right), \quad t = 0, 1, 2, \ldots$$

Теорема 2.2 *Пусть $f(\Omega)$ — ограниченная сверху, непрерывно дифференцируемая функция, и все Ω^t, начиная с некоторой итерации t^0 обладают свойствами:*

- $\forall j \in J \;\; \forall i \in I_j \;\; \omega_{ij}^t = 0 \to \omega_{ij}^{t+1} = 0$ *(сохранение нулей)*
- $\exists \epsilon > 0 \;\; \forall j \in J \;\; \forall i \in I_j \;\; \omega_{ij}^t \notin (0, \epsilon)$ *(отделимость от нуля)*
- $\exists \delta > 0 \;\; \forall j \in J \;\; \exists i \in I_j \;\; \omega_{ij}^t \frac{\partial f(\Omega^t)}{\partial \omega_{ij}} \geqslant \delta$ *(невырожденность)*

Тогда $f(\Omega^{t+1}) > f(\Omega^t)$ и $\left|\omega_{ij}^{t+1} - \omega_{ij}^t\right| \to 0$ при $t \to \infty$.

Доказательство данной теоремы получено И. А. Ирхиным как обобщение результатов о сходимости EM-алгоритма в тематическом моделировании [26].

3 Тематическое моделирование

Пусть D — конечное множество (коллекция) текстовых документов, W — конечное множество (словарь) всех употребляемых в них термов. *Термами* могут быть слова, нормальные формы слов, n-граммы или словосочетания, в зависимости от вида предварительной обработки текста. Каждый документ $d \in D$ представляет собой последовательность n_d термов $w_1, \ldots, w_{n_d}$ из словаря W.

Предположим, что порядок термов в документах не важен для определения его тематики. Это предположение называют гипотезой «мешка слов» (bag of words). Тогда документ d представляется подмножеством термов w, каждый из которых встречается n_{dw} раз.

Предположим, что появление термов в документе d зависит от тем, но не зависит от самого документа: $p(w\,|\,d,t) = p(w\,|\,t)$. Это предположение называют гипотезой условной независимости.

Согласно формуле полной вероятности, распределение термов в документе $p(w\,|\,d)$ описывается вероятностной смесью распределений термов в темах $\varphi_{wt} = p(w\,|\,t)$ с весами $\theta_{td} = p(t\,|\,d)$:

$$p(w\,|\,d) = \sum_{t\in T} p(w\,|\,t)\,p(t\,|\,d) = \sum_{t\in T} \varphi_{wt}\theta_{td}. \tag{3}$$

Вероятностная модель (3) описывает процесс порождения текстовой коллекции по известным распределениям $p(w\,|\,t)$ и $p(t\,|\,d)$.

Задача вероятностного тематического моделирования (PTM) — это обратная задача: по заданной коллекции D требуется найти параметры φ_{wt} и θ_{td}, при которых тематическая модель (3) хорошо приближает входные данные — частотные оценки условных вероятностей $\hat{p}(w\,|\,d) = n_{dw}/n_d$.

Параметры тематической модели — *матрица терминов тем* $\Phi = (\varphi_{wt})_{W\times T}$ и *матрица тем документов* $\Theta = (\theta_{td})_{T\times D}$ — оцениваются по критерию максимума логарифма правдоподобия:

$$\sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{t\in T} \varphi_{wt}\theta_{td} \;\to\; \max_{\Phi,\Theta} \tag{4}$$

при ограничениях неотрицательности и нормировки:

$$\sum_{w\in W} \varphi_{wt} = 1, \quad \varphi_{wt} \geqslant 0; \qquad \sum_{t\in T} \theta_{td} = 1, \quad \theta_{td} \geqslant 0. \tag{5}$$

Для лучшего понимания тематического моделирования полезно рассмотреть постановку задачи с четырёх точек зрения.

Во-первых, это задача приближённого низкорангового матричного разложения. Его ранг $|T|$, как правило, много меньше входных размерностей $|D|$ и $|W|$. Задача некорректно поставлена, поскольку её решение не единственно: $\Phi\Theta = (\Phi S)(S^{-1}\Theta)$ для невырожденных матриц S. Чтобы доопределить решение и сделать его устойчивым, к основному критерию необходимо добавлять регуляризаторы, учитывающие дополнительные прикладные знания или данные.

Во-вторых, это автокодировщик текстовых документов. Кодировщик $f_\Phi\colon \frac{n_{dw}}{n_d} \to \theta_d$ преобразует разреженное векторное представление документа $\hat{p}(w\,|\,d)$ размерности $|W|$ в тематическое векторное представление $\theta_d = p(t\,|\,d)$ размерности $|T|$. Линейный декодировщик $g_\Phi\colon \theta_d \to \Phi\theta_d$ пытается реконструировать исходное представление как можно точнее. Матрица Φ является параметром как кодировщика, так и декодировщика. Матрица $\Theta = (\theta_1, \ldots, \theta_D)$ является результатом кодирования всех документов коллекции. Это важное различие ролей двух матриц ускользает от внимания, если рассматривать тематическую модель только с точки зрения матричного разложения.

В-третьих, это способ мягкой би-кластеризации документов по множеству тематических кластеров T. Каждый документ d и каждый терм w не относится жёстко к одному кластеру, а мягко аллоцируется по всем кластерам согласно распределениям $p(t\,|\,d)$ и $p(t\,|\,w)$. Модель

позволяет также оценивать тематические распределения для терма в документе $p(t\,|\,d,w)$, для предложения $p(t\,|\,s)$ или любого текстового фрагмента. Вообще, распределение вида $p(t\,|\,x)$ будем называть тематическим векторным представлением, тематическим эмбедингом или *тематикой* объекта x.

В-четвёртых, это языковая модель, которая предсказывает появление слов в документах. Следует признать, что в данном качестве обычные тематические модели довольно слабы. Хорошо предсказать слово возможно только при тщательном анализе его локального контекста, который разрушается гипотезой «мешка слов». В тематическом моделировании изобретено довольно много способов отказаться от этой гипотезы и обрабатывать текст как последовательность термов. Но есть и другой, неустранимый, недостаток: вряд ли стоит ожидать, что появление слова определяется только тематикой, даже если она определяется по локальному контексту. Глубокие нейронные сети на основе моделей внимания [4] и архитектуры трансформера, такие как BERT [16] и GPT-3 [33] моделируют весь комплекс языковых явлений и предсказывают слова в тексте лучше, чем это делают люди. Однако эти модели неинтерпретируемы: мы не можем знать, какие именно языковые явления и каким образом смоделировала нейросеть. Также неясно, какой смысл имеет каждая координата векторного представления текста.

Тематические эмбединги, в отличие от нейросетевых, обладают свойством интерпретируемости. Тема может рассказать о себе наиболее частотными словами из распределения $p(w\,|\,t)$, либо целыми фразами, отобранными из текстов методами экстрактивной суммаризации или автоматического именования тем [35]. Также и любой тематический эмбединг $p(t\,|\,x)$ способен рассказать о себе словами или фразами естественного языка.

Целью тематического моделирования является не столько предсказание слов в документах, сколько выявление тематической структуры текстовой коллекции, определение тематики документов и связанных с ними объектов, объяснение тем на естественном языке.

4 Аддитивная регуляризация

Для решения некорректно поставленной задачи стохастического матричного разложения добавим к логарифму правдоподобия (4) регуляризационный критерий $R(\Phi,\Theta)$, при прежних ограничениях неотрицательности и нормировки (5):

$$L(\Phi,\Theta)=\sum_{d\in D}\sum_{w\in d} n_{dw}\ln\sum_{t\in T}\varphi_{wt}\theta_{td}+R(\Phi,\Theta)\;\to\;\max_{\Phi,\Theta}. \qquad (6)$$

В общем случае к тематической модели может предъявляться несколько требований, каждому из которых соответствует свой регуляризатор $R_i(\Phi,\Theta)$. Метод скаляризации критериев для многокритериальной оптимизации приводит к подходу *аддитивной регуляри-*

зации тематических моделей (ARTM), предложенному в [52]:

$$R(\Phi,\Theta) = \sum_{i=1}^{k} \tau_i R_i(\Phi,\Theta),$$

где неотрицательные коэффициенты регуляризации τ_i, $i = 1,\dots,k$, являются гиперпараметрами алгоритма обучения.

Теорема 4.1 *Пусть функция $R(\Phi,\Theta)$ непрерывно дифференцируема. Тогда точка (Φ,Θ) локального экстремума задачи (6) с ограничениями (5) удовлетворяет системе уравнений со вспомогательными переменными $p_{tdw} = p(t\,|\,d,w)$, если из решения исключить нулевые столбцы матриц Φ, Θ:*

$$p_{tdw} = \underset{t\in T}{\operatorname{norm}}\big(\varphi_{wt}\theta_{td}\big); \tag{7}$$

$$\varphi_{wt} = \underset{w\in W}{\operatorname{norm}}\left(n_{wt} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}}\right); \qquad n_{wt} = \sum_{d\in D} n_{dw} p_{tdw}; \tag{8}$$

$$\theta_{td} = \underset{t\in T}{\operatorname{norm}}\left(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right); \qquad n_{td} = \sum_{w\in d} n_{dw} p_{tdw}. \tag{9}$$

Доказательство можно найти в [55], но намного проще вывести его из теоремы 2.1. Перепишем (7) в следующем виде:

$$p_{tdw} = \underset{t\in T}{\operatorname{norm}}\big(\varphi_{wt}\theta_{td}\big) = \frac{\varphi_{wt}\theta_{td}}{\sum_s \varphi_{ws}\theta_{sd}} = \frac{\varphi_{wt}\theta_{td}}{p(w\,|\,d)}.$$

Применим к задаче (6) формулу (1) и выделим в полученных выражениях вспомогательные переменные p_{tdw}:

$$\begin{aligned}
\varphi_{wt} &= \underset{w\in W}{\operatorname{norm}}\left(\varphi_{wt}\frac{\partial L}{\partial \varphi_{wt}}\right) = \underset{w\in W}{\operatorname{norm}}\left(\varphi_{wt}\sum_{d\in D}\frac{n_{dw}\theta_{td}}{p(w\,|\,d)} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}}\right) \\
&= \underset{w\in W}{\operatorname{norm}}\left(\sum_{d\in D} n_{dw}p_{tdw} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}}\right); \\
\theta_{td} &= \underset{t\in T}{\operatorname{norm}}\left(\theta_{td}\frac{\partial L}{\partial \theta_{td}}\right) = \underset{t\in T}{\operatorname{norm}}\left(\theta_{td}\sum_{w\in d}\frac{n_{dw}\varphi_{wt}}{p(w\,|\,d)} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right) \\
&= \underset{t\in T}{\operatorname{norm}}\left(\sum_{w\in d} n_{dw}p_{tdw} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right).
\end{aligned}$$

Нулевые столбцы в матрицах Φ и Θ получаются в тех случаях, когда не выполнено условие теоремы 2.1 о положительности хотя бы одной координаты в нормируемом векторе. Такие вектор-столбцы удаляются из матриц, что разрешено условием теоремы.

Теорема доказана.

Тема t *вырождена*, если $n_{wt} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}} \leqslant 0$ для всех $w \in W$.

Вырожденность темы является следствием чрезмерно сильного разреживающего воздействия регуляризатора R. Обнуление столбца матрицы Φ означает, что модели стало выгодно вообще не использовать данную тему. Сокращение числа тем может быть желательным побочным эффектом регуляризации.

Документ d *вырожден*, если $n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}} \leqslant 0$ для всех $t \in T$.

Вырожденность документа означает, что модель не в состоянии его описать, например, если он слишком короткий или вовсе не соответствует тематике коллекции.

Вырожденность значительного числа тем или документов может свидетельствовать об избыточной регуляризации и служить сигналом для уменьшения некоторых коэффициентов регуляризации.

Для обучения тематической модели система (7)–(9) решается численно, методом простых итераций. Это приводит к алгоритму Expectation–Maximization, в котором на каждой итерации выполняются два шага: *E-шаг* (7) и *M-шаг* (8)–(9). При рациональной реализации этого алгоритма каждая итерация выполняется за один линейный проход по коллекции. Для каждого терма w в каждом документе d тематический вектор $p(t|d,w)$ вычисляется по формуле E-шага, используется для обновления счётчиков n_{wt} и n_{td}, и сразу забывается, чтобы минимизировать расход памяти. Быстрый онлайновый алгоритм, реализованный в библиотеке `BigARTM` [53], использует распараллеливание, разбиение коллекции на батчи, управление частотой обновления матрицы Φ и ещё несколько приёмов для увеличения скорости вычислений [20, 3]. В результате `BigARTM` обучает тематические модели во много раз быстрее других свободно доступных инструментов, таких как Gensim и Vowpal Wabbit, на некоторых задачах опережая их в 20 раз [31].

Модель вероятностного латентного семантического анализа (PLSA) исторически является первой вероятностной тематической моделью [22]. В ARTM ей соответствует нулевой регуляризатор

$$R(\Phi, \Theta) = 0.$$

Модель латентного размещения Дирихле (LDA) [8] — это наиболее известная байесовская модель, в которой на столбцы матриц Φ и Θ накладываются ограничения в виде априорных распределений Дирихле. В ARTM ей соответствует кросс-энтропийный регуляризатор [31]

$$R(\Phi, \Theta) = \sum_{t \in T} \sum_{w \in W} \beta_w \ln \varphi_{wt} + \sum_{d \in D} \sum_{t \in T} \alpha_t \ln \theta_{td}. \tag{10}$$

Если значения гиперпараметров β_w, α_t положительны, то регуляризация сглаживает условные распределения φ_{wt}, θ_{td} приближая их к заданным векторам $\underset{w}{\text{norm}}(\beta_w)$, $\underset{t}{\text{norm}}(\alpha_t)$. Если же β_w, α_t отрицательны, то воздействие регуляризатора противоположно, вместо сглаживания происходит разреживание, как видно из формул M-шага:

$$\varphi_{wt} = \underset{w \in W}{\text{norm}}\big(n_{wt} + \beta_w\big); \qquad \theta_{td} = \underset{t \in T}{\text{norm}}\big(n_{td} + \alpha_t\big).$$

В байесовской интерпретации гиперпараметры ограничены снизу: $\beta_w > -1$, $\alpha_t > -1$, из-за свойств распределения Дирихле. Поэтому разреживающее воздействие оказывается слабым. В оптимизационной интерпретации ARTM таких ограничений нет, поскольку априорные распределения Дирихле в модель не вводятся.

5 Сравнение с байесовским обучением

Пусть для общности X — наблюдаемая выборка данных (в нашем случае это коллекция текстовых документов), $p(X\,|\,\Omega)$ — вероятностная модель данных с параметрами Ω (в нашем случае это матрицы Φ, Θ), $p(\Omega\,|\,\gamma)$ — *априорное распределение* в пространстве параметров модели, имеющее гиперпараметры γ (в модели LDA это распределения Дирихле с гиперпараметрами β_w, α_t). Тогда *апостериорное распределение* параметров, согласно формуле Байеса, имеет вид

$$p(\Omega\,|\,X,\gamma) = \frac{p(\Omega, X\,|\,\gamma)}{p(X\,|\,\gamma)} \propto p(X\,|\,\Omega)\,p(\Omega\,|\,\gamma),$$

где символ $\propto$ означает «равно с точностью до нормировки». Байесовский вывод полезен в тех задачах анализа данных, где апостериорные распределения используются для получения интервальных оценок, проверки статистических гипотез о параметрах модели, сэмплирования параметров и т.д. Однако в практике тематического моделирования байесовский вывод производится исключительно ради получения точечной оценки параметров Ω:

$$\Omega := \arg\max_{\Omega} p(\Omega\,|\,X,\gamma).$$

Максимизация апостериорной вероятности (MAP) даёт точечную оценку Ω, минуя промежуточный шаг вывода апостериорного распределения, который часто оказывается приближённым и трудоёмким:

$$\Omega := \arg\max_{\Omega}\big(\ln p(X\,|\,\Omega) + \ln p(\Omega\,|\,\gamma)\big).$$

В результате дебайесизации логарифм априорного распределения становится критерием регуляризации $R(\Omega) = \ln p(\Omega\,|\,\gamma)$. Его можно отделить от конкретной модели и добавлять к другим моделям.

Аддитивная регуляризация обобщает MAP на любые регуляризаторы, в том числе не имеющие вероятностной природы, а также их линейные комбинации (что не ухудшает свойства сходимости):

$$\Omega := \arg\max_{\Omega}\big(\ln p(X\,|\,\Omega) + \sum_{i=1} \tau_i R_i(\Omega)\big).$$

К недостаткам байесовского вывода можно отнести техническую сложность добавления и комбинирования требований к модели как оптимизационных критериев. В нём нет удобных механизмов регуляризации, поскольку нет, собственно, и задачи оптимизации по Ω. Дополнительная информация привносится либо через априорные распределения, либо в саму структуру модели. Если априорные распределения не являются распределениями Дирихле, вывод заметно усложняется.

Распределение Дирихле играет особую роль в байесовском тематическом моделировании. Оно не имеет убедительных лингвистических обоснований, тем не менее, в литературе большинство моделей строятся с его использованием. Это объясняется исключительно математическим удобством сопряжённости распределений Дирихле с мультиномиальным распределением. В ARTM нет никаких оснований предпочитать распределение Дирихле другим регуляризаторам.

Аддитивность регуляризаторов приводит к модульной технологии тематического моделирования, которая реализована в проекте `BigARTM` [53]. При решении прикладных задач комбинирование готовых регуляризаторов позволяет строить модели с заданными свойствами без математических выкладок и программирования. Создание такой технологии в рамках байесовского подхода едва ли возможно.

6 Обзор моделей и регуляризаторов

В статье [31] показана дебайесизация многих тематических моделей, исходно сформулированных в рамках байесовской парадигмы.

Сочетание регуляризаторов сглаживания, разреживания и декоррелирования хорошо зарекомендовало себя на практике во многих исследованиях [54, 55, 59]. Регуляризатор декоррелирования тем

$$R(\Phi) = -\frac{\tau}{2}\sum_{t\in T}\sum_{s\in T\setminus t}\sum_{w\in W}\varphi_{wt}\varphi_{ws}$$

не только делает темы различными, но и способствует выделению общеупотребительных слов в отдельные «фоновые» темы и очищает от них все остальные темы [48].

Модели с частичным обучением используют регуляризатор сглаживания тем как столбцов матрицы Φ, чтобы задать часть лексики, вокруг которой образуются нужные предметные темы. Эта техника использовалась для поиска редкой информации в социальных медиа, связанной с болезнями, симптомами и методами лечения [40, 41], с преступностью и экстремизмом [34, 46], с национальностями и межнациональными отношениями [9, 32, 38]. Например, для поиска заданного числа этно-релевантных тем в рамках ARTM применялось сглаживание по словарю этнонимов; после этого тематическая модель сама определяла, как темы специализируются по этничностям [1, 2]. В том числе, образовывались полиэтничные темы, помогавшие социологам выявлять особенности межэтнических отношений.

Мультимодальная тематическая модель описывает документы, содержащие наряду со словами термы других модальностей: категории, авторы, время, теги, названия, пользователи, и т.д. Для каждой модальности $m \in M$ вводится свой словарь термов W^m, своя матрица Φ^m с нормированными столбцами, свой критерий логарифма правдоподобия. Максимизируется взвешенная сумма этих критериев:

$$\sum_{m\in M}\tau_m\sum_{d\in D}\sum_{w\in W^m} n_{dw}\ln\sum_{t\in T}\varphi^m_{wt}\theta_{td} + R\big(\{\Phi^m\},\Theta\big) \;\to\; \max_{\Phi,\Theta}. \quad (11)$$

Мультимодальные данные помогают более точно определять тематику документов, и, наоборот, тематическая модель может использоваться для выявления семантики модальностей или предсказания пропущенных метаданных.

Тематическая модель классификации является частным случаем мультимодальной, с модальностью C классов или категорий. Тематическая модель модальности классов может непосредственно использоваться в качестве линейной вероятностной модели классификации

документов d, при этом в роли вектора признаков выступает тематический вектор документа:

$$p(c\,|\,d) = \sum_{t\in T} p(c\,|\,t)p(t\,|\,d) = \sum_{t\in T} \varphi_{ct}\theta_{td}.$$

Эксперименты в [45] показали, что тематические модели превосходят обычные методы многоклассовой классификации на больших текстовых коллекциях с большим числом несбалансированных, пересекающихся, взаимозависимых классов. В [51] те же выводы на тех же коллекциях были воспроизведены для мультимодальной ARTM. *Несбалансированность* означает, что классы могут содержать как малое, так и очень большое число документов. В случае *пересекающихся* классов документ может относиться к одному или нескольким классам. *Взаимозависимые* классы имеют общие термы и темы, поэтому при классификации документа могут вступать в конкуренцию.

Мультиязычная тематическая модель — также частный случай мультимодальной, когда в роли модальностей выступают языки. Оказалось, что связывания параллельных текстов в общий документ достаточно для синхронизации тем в двух языках и кросс-язычного поиска [56]. Регуляризаторы на основе двуязычных словарей были предложены в [17], однако оказалось, что основной вклад в качество поиска даёт всё-таки связывание текста с его переводами.

Модель трёхматричного разложения связана с предположением, что темы порождаются не документом, а одной из модальностей, например, категориями или авторами. Например, для автор-тематической модели (author-topic model, ATM) [44],

$$p(w\,|\,d) = \sum_{t\in T} p(w\,|\,t) \sum_{a\in A} p(t\,|\,a)p(a\,|\,d) = \sum_{t\in T}\sum_{a\in A} \varphi_{wt}\psi_{ta}\pi_{ad},$$

где A — словарь авторов. Приведённый в [31] регуляризованный EM-алгоритм для этой модели легко получается как следствие из теоремы о максимизации на единичных симплексах 2.1.

Иерархические тематические модели делят темы на всё более мелкие подтемы и отличаются разнообразием подходов и способов оценивания иерархий [60]. Нисходящая стратегия на основе ARTM предложена в [14] и улучшена в [6]. Иерархия строится по уровням сверху вниз, на каждом следующем дочернем уровне число тем больше, чем на предыдущем родительском. Каждый уровень представляет собой обычную «плоскую» тематическую модель, которая связывается с родительским уровнем условными вероятностями $\psi_{st} = p(s\,|\,t)$ подтем $s \in S$ в родительских темах $t \in T$. Регуляризатор требует, чтобы родительские темы φ_{wt} аппроксимировались вероятностной смесью дочерних тем φ_{ws} с коэффициентами ψ_{st}:

$$R(\Phi,\Psi) = \tau \sum_{t\in T}\sum_{w\in W} n_{wt} \ln \sum_{s\in S} \varphi_{ws}\psi_{st}. \tag{12}$$

Задача максимизации $R(\Phi,\Psi)$ с точностью до обозначений совпадает с основной задачей тематического моделирования (4), если считать родительские темы t *псевдодокументами* с частотами термов

$n_{wt} = n_t \varphi_{wt}$. Данный регуляризатор реализуется простым добавлением $|T|$ псевдодокументов в коллекцию перед построением каждого дочернего уровня. Матрица связей Ψ получается на выходе модели в столбцах матрицы Θ, соответствующих псевдодокументам.

Мультимодальные иерархические тематические модели хорошо показали себя в задачах тематического поиска документов по документам [23, 24]. Комбинирование регуляризаторов декоррелирования, разреживания и сглаживания вместе с модальностями n-грамм, авторов и категорий значимо улучшает качество поиска и позволяет достичь точности и полноты выше 90%. Оптимальная (с точки зрения качества поиска) размерность тематических векторов на третьем уровне иерархии оказалась в несколько раз выше, чем у плоской модели. Это означает, что при постепенном дроблении тем тематические векторы накапливают в себе больше полезной информации.

Тематическая модель мнений фактически является двухуровневой иерархией, в которой верхний уровень строится обычным образом по словарю термов и выделяет событийные темы в новостном потоке [19]. Второй уровень строится по другим модальностям, что позволяет разделить тему не на подтемы, а на поляризованные мнения о событии. В качестве таких модальностей предлагается брать именованные сущности с позитивными и негативными тональностями, именованные сущности с их семантическими ролями, триплеты «субъект, предикат, объект». Эксперименты показали, что каждая из трёх модальностей важна для повышения качества кластеризации мнений. Аналогичная двухуровневая иерархия была предложена в [42], где синтаксические модальности использовались для разделения тем верхнего уровня на клиентские интенты по коллекции диалогов контактного центра.

Стратегии оптимизации гиперпараметров. Аддитивная регуляризация проигрывает байесовскому моделированию только в одном аспекте. Чем больше используется регуляризаторов, тем больше коэффициентов регуляризации приходится подбирать, тем более тщательной балансировки они требуют. Первые исследования показали, что регуляризаторы могут даже мешать друг другу, а понимание их взаимодействий приводит к стратегии последовательного добавления регуляризаторов в модель [54].

Апробированная последовательность «декоррелирование, разреживание Θ, сглаживание Φ» использовалась в последующих работах по тематическому поиску [59, 23]. Позже к ним был добавлен перебор весов модальностей, весов псевдодокументов в иерархической модели, подбор числа тем на каждом уровне иерархии [24]. Каждый регуляризатор включается, начиная с определённой итерации, что требует многократного перестроения модели с промежуточной стартовой точки, но при разных значениях коэффициента регуляризации. Эти значения вполне достаточно перебрать по грубой сетке, чтобы определить рабочий диапазон коэффициента и точку устойчивого максимума. Сходимость итерационного процесса и качество модели по совокупности критериев оцениваются визуально.

Позже эта методика была расширена и положена в основу верхнеуровневой библиотеки TopicNet, которая использует BigARTM, но

скрывает от пользователя технические детали [11]. Пользователь задаёт только общую стратегию регуляризации — в каком порядке включать регуляризаторы. TopicNet автоматизирует проведение серий вычислительных экспериментов по перебору регуляризованных моделей, обеспечивая журнализацию и визуализацию.

Ещё более общий метод автоматической настройки гиперпараметров для ARTM предложен в [30] на основе эволюционного алгоритма и представления процесса обучения как многоэтапной стратегии изменения гиперпараметров. В работе [29] этот метод дополнен суррогатной моделью оценивания качества, что позволило сократить время поиска гиперпараметров.

7 Гиперграфовые тематические модели транзакционных данных

Тематические модели текстовых коллекций описывают вхождения слов в документы. Мультимодальные модели описывают документы, в которых содержатся термы различных модальностей: слова, теги, категории, авторы, и т. д. Во всех этих случаях модель описывает парные взаимодействия между документами и термами. В более сложных приложениях исходные данные могут описывать транзакции между тремя и более объектами различных модальностей. Например, в сети интернет-рекламы «пользователь u кликнул объявление b на странице s»; в социальной сети «пользователь u написал слово w на странице блога d»; в сети продаж «покупатель b купил у продавца s товар g»; в пассажирских авиаперевозках «клиент u вылетел из аэропорта x в аэропорт y самолётом авиакомпании a»; в рекомендательной системе «клиент u оценил фильм f в ситуативном контексте s». Ещё одной модальностью может быть время транзакции. Во всех приведённых выше примерах транзакция нескольких объектов не сводится к их парным взаимодействиям.

Для моделирования транзакционных данных удобно использовать гиперграф $\Gamma = \langle V, E \rangle$, который определяется множеством вершин-термов V и множеством рёбер-транзакций E. Каждое ребро e из E является подмножеством из двух или более вершин, $e \subset V$. Задача заключается в том, чтобы по наблюдаемой выборке транзакций восстановить неизвестные тематические распределения вершин $p(t|v)$.

Каждая вершина имеет модальность m из множества M. Обозначим через V_m множество вершин модальности m. В обычных тематических моделях есть только две модальности: документы $V_1 = D$ и термы $V_2 = W$; каждая транзакция представляется ребром из двух вершин $e = (d, w)$ и описывает вхождение терма w в документ d. При этом гиперграф является двудольным графом.

В более сложных приложениях транзакции могут иметь различные типы. Например, в сети интернет-рекламы, кроме данных типа (u, b, s) о кликах пользователей u по объявлениям b на страницах s, могут иметься данные о посещениях страниц пользователями (u, s),

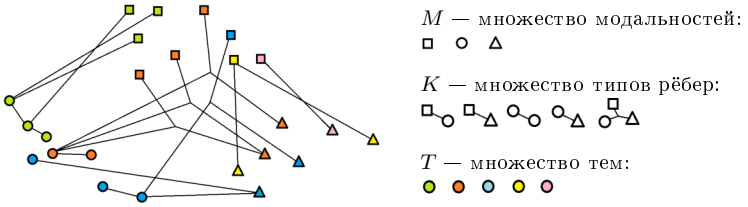

Рис. 1: Пример гиперграфа с вершинами трёх модальностей, рёбрами-транзакциями пяти типов и пятью темами.

о содержании термов w в текстах объявлений (b, w), страниц (s, w) и запросов пользователей (u, w).

Пусть задано множество типов транзакций K. *Транзакционные данные* типа k — это выборка рёбер $E_k \subset E$. Каждое ребро $e \in E_k$ входит в выборку n_{ke} раз, и с каждым вхождением ребра связана своя латентная тема $t \in T$. На рис. 1 показан пример гиперграфа.

Будем полагать, что в каждой транзакции $e \in E$ имеется одна выделенная вершина d, называемая *контейнером*, и обозначать ребро через $e = (d, x)$, где x — множество всех остальных вершин ребра. Аналогично документу, с контейнером связано распределение тем $p(t|d)$. Множество всех вершин-контейнеров обозначим через D.

Примем несколько гипотез условной независимости. Предположим, что ни распределения тем $p(t|d)$ в контейнере d, ни распределения вершин в темах $p(v|t)$ не зависят от типа ребра k. Далее предположим, что процесс порождения ребра $(d, x) \in E_k$ состоит из двух шагов. Сначала порождается тема t из распределения $p(t|d)$. Затем порождается множество вершин $x \subset V$, причём каждая вершина $v \in x$ модальности m порождается независимо от остальных вершин ребра, из своего распределения $p(v|t)$ над множеством V_m.

Тематическая модель выражает вероятности появления рёбер гиперграфа через условные распределения, связанные с их вершинами:

$$p(x|d) = \sum_{t \in T} p(t|d) \prod_{v \in x} p(v|t) = \sum_{t \in T} \theta_{td} \prod_{v \in x} \varphi_{vt}.$$

В матричных обозначениях параметрами модели являются матрицы Θ и Φ_m, $m \in M$, как в мультимодальной тематической модели (11).

Для обучения модели максимизируется сумма логарифмов правдоподобия по типам рёбер k с весами τ_k и регуляризатором $R(\Phi, \Theta)$, при обычных ограничениях неотрицательности и нормировки:

$$\sum_{k \in K} \tau_k \sum_{dx \in E_k} n_{kdx} \ln \left(\sum_{t \in T} \theta_{td} \prod_{v \in x} \varphi_{vt} \right) + R(\Phi, \Theta) \to \max_{\Phi, \Theta}; \qquad (13)$$

$$\sum_{v \in V_m} \varphi_{vt} = 1, \ \varphi_{vt} \geqslant 0; \qquad \sum_{t \in T} \theta_{td} = 1, \ \theta_{td} \geqslant 0.$$

Теорема 7.1 *Пусть функция $R(\Phi, \Theta)$ непрерывно дифференцируема. Точка локального максимума (Φ, Θ) задачи (13) удовлетворяет*

системе уравнений относительно параметров модели φ_{vt}, θ_{td} *и вспомогательных переменных* $p_{tdx} = p(t\,|\,d,x)$, *если из решения исключить нулевые столбцы матриц* Φ_m, Θ:

$$p_{tdx} = \underset{t\in T}{\operatorname{norm}}\Big(\theta_{td}\prod_{v\in x}\varphi_{vt}\Big); \tag{14}$$

$$\varphi_{vt} = \underset{v\in V_m}{\operatorname{norm}}\Big(\sum_{k\in K}\sum_{dx\in E_k}[v\in x]\,\tau_k n_{kdx}p_{tdx} + \varphi_{vt}\frac{\partial R}{\partial\varphi_{vt}}\Big); \tag{15}$$

$$\theta_{td} = \underset{t\in T}{\operatorname{norm}}\Big(\sum_{k\in K}\sum_{dx\in E_k}\tau_k n_{kdx}p_{tdx} + \theta_{td}\frac{\partial R}{\partial\theta_{td}}\Big); \tag{16}$$

Доказательство. Воспользуемся теоремой 2.1 о максимизации на единичных симплексах, выделив вспомогательные переменные p_{tdx}, определённые в (14):

$$\begin{aligned}
\varphi_{vt} &= \underset{v\in V_m}{\operatorname{norm}}\Big(\varphi_{vt}\sum_{k\in K}\tau_k\sum_{dx\in E_k}n_{kdx}\frac{\theta_{td}}{p(x\,|\,d)}\frac{\partial}{\partial\varphi_{vt}}\prod_{u\in x}\varphi_{ut} + \varphi_{vt}\frac{\partial R}{\partial\varphi_{vt}}\Big)\\
&= \underset{v\in V_m}{\operatorname{norm}}\Big(\sum_{k\in K}\sum_{dx\in E_k}\tau_k n_{kdx}[v\in x]p_{tdx} + \varphi_{vt}\frac{\partial R}{\partial\varphi_{vt}}\Big);\\
\theta_{td} &= \underset{t\in T}{\operatorname{norm}}\Big(\theta_{td}\sum_{k\in K}\tau_k\sum_{x\in d}n_{kdx}\frac{1}{p(x\,|\,d)}\prod_{v\in x}\varphi_{vt} + \theta_{td}\frac{\partial R}{\partial\theta_{td}}\Big)\\
&= \underset{t\in T}{\operatorname{norm}}\Big(\sum_{k\in K}\sum_{x\in d}\tau_k n_{kdx}p_{tdx} + \theta_{td}\frac{\partial R}{\partial\theta_{td}}\Big).
\end{aligned}$$

Теорема доказана.

Гиперграфовая модель является широким обобщением обычных тематических моделей. Несмотря на это, вывод EM-алгоритма для неё с помощью теоремы 2.1 оказался не сложнее, чем в случае обычного матричного разложения. Данный алгоритм реализован в `BigARTM`.

8 Гиперграфовые тематические модели для рекомендательных систем

Пусть U — конечное множество пользователей, I — конечное множество объектов, которые пользователи могут выбирать. Вероятностная тематическая модель предсказывает предпочтения пользователей:

$$p(i\,|\,u) = \sum_{t\in T}p(i\,|\,t)\,p(t\,|\,u).$$

Она эквивалентна тематической модели текстовой коллекции с точностью до терминологии: «документы → клиенты», «термы → объекты», «темы → интересы». Исходными данными являются счётчики n_{ui} транзакций пользователя u с объектом i. В зависимости от приложения это могут быть покупки, просмотры, лайки и т. д.

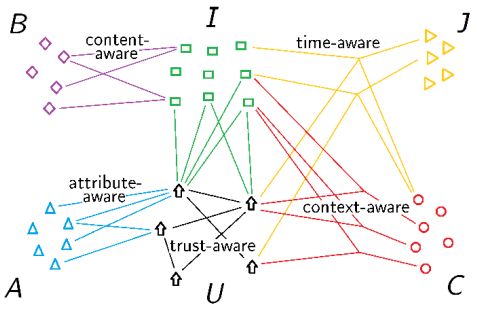

Рис. 2: Типы транзакций между шестью модальностями в рекомендательной системе: клиенты U, объекты I, атрибуты клиентов A, свойства объектов B, ситуативные контексты C, интервалы времени J.

В рекомендательных системах существует проблема «холодного старта»: новому клиенту нечего порекомендовать, поскольку нет истории его предпочтений; также и новый товар некому порекомендовать, поскольку его ещё никто не выбирал. Для решения этой проблемы привлекаются дополнительные данные о клиентах и объектах. В частности, это могут быть данные n_{ua} о наличии у клиента u атрибута $a \in A$ или данные n_{ib} о наличии у объекта i свойства $b \in B$. Если объекты имеют текстовые описания, то B — это словарь термов, используемых в этих описаниях. Такие рекомендательные системы называются, соответственно, учитывающими атрибуты (attribute-aware) и учитывающими контент (content-aware).

В качестве дополнительных данных могут также использоваться советы клиентов друг другу. Это попарные взаимодействия между клиентами $n_{uu'}$ или данные о доверии (trust-aware).

Предпочтения клиентов могут изменяться со временем или зависеть от ситуации. Для учёта такой информации вводятся ещё две модальности: множество ситуаций C и множество интервалов времени J. Взаимодействия между ними описываются транзакциями из трёх или более термов, например: n_{uic} — клиент u выбрал объект i в ситуации c; n_{uicj} — клиент u выбрал объект i в ситуации c в интервале времени j. Такие системы называются, соответственно, учитывающими контекст (context-aware) и учитывающими время (time-aware).

Перечисленные типы моделей в литературе вводились по отдельности [13]. Гиперграфовая модель позволяет объединить их для получения векторных представлений термов любой природы в общем семантическом векторном пространстве, рис. 2.

Данные рекомендательных систем отличаются от текстовых коллекций тем, что в них нет естественного аналога документа или контейнера. Документы статичны и не меняются во времени, тогда как множество транзакций (u, i) увеличивается со временем как для пользователя u, так и для объекта i.

Будем считать, что в рёбрах гиперграфа $x \subset V$ нет никакой выделенной вершины-контейнера. Для генерации ребра сначала порождается тема t из распределения $\pi_t = p(t)$, общего для всей коллекции;

затем вершины $v \in x$ порождаются независимо друг от друга из распределений $\varphi_{vt} = p(v\,|\,t)$ над модальностями V_m:

$$p(x) = \sum_{t\in T} p(t) \prod_{v\in x} p(v\,|\,t) = \sum_{t\in T} \pi_t \prod_{v\in x} \varphi_{vt}.$$

Тематические модели, в которых документы выступают в роли одной из модальностей, называются симметричными [50]. Соответственно записывается задача максимизации регуляризованного правдоподобия при ограничениях нормировки и неотрицательности:

$$\sum_{k\in K} \tau_k \sum_{x\in E_k} n_{kx} \ln\Bigg(\sum_{t\in T} \pi_t \prod_{v\in x} \varphi_{vt}\Bigg) \;+\; R(\Phi,\pi) \to \max_{\Phi,\pi}; \tag{17}$$

$$\sum_{v\in V_m} \varphi_{vt} = 1,\; \varphi_{vt} \geqslant 0; \qquad \sum_{t\in T} \pi_t = 1,\; \pi_t \geqslant 0.$$

Теорема 8.1 *Пусть функция $R(\Phi,\pi)$ непрерывно дифференцируема. Точка локального максимума (Φ,π) задачи (17) удовлетворяет системе уравнений относительно параметров модели φ_{vt}, π_t и вспомогательных переменных $p_{tx} = p(t\,|\,x)$, если из решения исключить нулевые столбцы матриц Φ_m:*

$$p_{tx} = \underset{t\in T}{\operatorname{norm}}\Bigg(\pi_t \prod_{v\in x} \varphi_{vt}\Bigg). \tag{18}$$

$$\varphi_{vt} = \underset{v\in V_m}{\operatorname{norm}}\Bigg(\sum_{k\in K} \sum_{x\in E_k} [v\in x]\,\tau_k n_{kx} p_{tx} + \varphi_{vt}\frac{\partial R}{\partial \varphi_{vt}}\Bigg); \tag{19}$$

$$\pi_t = \underset{t\in T}{\operatorname{norm}}\Bigg(\sum_{k\in K} \sum_{x\in E_k} \tau_k n_{kx} p_{tx} + \pi_t\frac{\partial R}{\partial \pi_t}\Bigg); \tag{20}$$

Доказательство, как и в случае предыдущей теоремы, проводится по теореме о максимизации на единичных симплексах.

В `BigARTM` такая модель не реализована, но её нетрудно симулировать. Коллекция разбивается некоторым образом на документы (например, по времени транзакций), и вводится сильный регуляризатор сглаживания столбцов матрицы Θ в сторону общего вектора (n_t), просуммированного по всем документам.

9 Модели последовательного текста

Гипотеза мешка слов является одним из наиболее критикуемых допущений в тематическом моделировании. Использование данных о сочетаемости слов или о порядке слов в тексте позволяет полностью или частичного уйти от этого предположения.

Тематические модели n-грамм эксплуатируют тот факт, что устойчивые сочетания из n подряд идущих слов часто, хотя и не всегда, являются специальными терминами или названиями. Они говорят о темах гораздо больше, чем те же слова, взятые по отдельности. Поэтому темы, построенные на словарях n-грамм, намного луч-

ше интерпретируются, чем построенные на униграммах [58, 27]. Существует два подхода к использованию n-грамм в тематическом моделировании. В первом случае словарь n-грамм строится на этапе препроцессинга текстов методами автоматического выделения терминов, ключевых слов или коллокаций [18]. Построенный словарь используется в качестве модальности. Второй подход более сложный, в нём обучение тематической модели объединяется с формированием словаря n-грамм [57, 58]. Концентрация распределения $p(t\,|\,w)$ в одной или нескольких темах является сильным признаком того, что n-грамма w является термином предметный области.

Тематическая модель сети слов предсказывает появление слова поблизости от другого слова, вместо того, чтобы предсказывать его в документе. Поблизости означает, например, в одном предложении или на расстоянии не более 10 слов. Для каждого слова $u \in W$ построим псевдо-документ d_u, состоящий из всех слов, встречающихся поблизости от слова u повсюду в коллекции. Обозначим через n_{uw} число вхождений слова w в псевдо-документ d_u.

Тематическая модель сети слов WNTM (word network topic model) [63] и более ранняя тематическая модель слов WTM (word topic model) [12] предсказывают появление слова в окрестности другого слова:

$$p(w\,|\,u) = \sum_{t\in T} p(w\,|\,t)\,p(t\,|\,d_u) = \sum_{t\in T} \varphi_{wt}\theta_{tu}.$$

Логарифм правдоподобия может служить регуляризатором для обычной модели документов, либо основным критерием обучения. В первом случае модель строится по исходной коллекции документов, аугментированной псевдодокументами слов; во втором случае используются только псевдодокументы:

$$\sum_{u,w\in W} n_{uw} \ln \sum_{t\in T} \varphi_{wt}\theta_{tu} \;\to\; \max_{\Phi,\Theta}.$$

Согласно *дистрибутивной гипотезе* (distributional hypothesis) смысл слова определяется распределением всех слов, в окружении которых оно встречается [21]. Слова, встречающиеся в схожих контекстах, имеют схожую семантику, и в модели они должны получать близкие векторы. Векторные представления слов, реализованные в программе `word2vec` [36, 37], также строятся по данным о парной сочетаемости слов. Известно, что для них выполняются векторные равенства на ассоциативных парах слов, например,

$$\text{король} - \text{королева} = \text{мужчина} - \text{женщина};$$

$$\text{Москва} - \text{Пекин} = \text{Россия} - \text{Китай}.$$

В [43] показано, что аддитивно регуляризованная WNTM также обладает этим свойством, в отличие от обычных тематических моделей. Более того, тематические векторы обладают также покоординатной интерпретируемостью, в отличие от word2vec и нейросетевых эмбедингов. Такие модели лучше подходят для анализа коротких текстов, таких как твиты или заголовки новостей.

Тематические модели предложений строятся как частный случай тематической модели гиперграфа. Вершинами гиперграфа являются слова, рёбрами — предложения. Такие модели предлагались и в байесовской парадигме, но в других терминах. Это тематическая модель предложений senLDA [5] и модель коротких сообщений Twitter-LDA [62]. В качестве транзакций можно брать не только предложения, но и любые словосочетания, синтагмы, именные группы, лексические цепочки, и вообще любые группы слов, относительно которых разумно предполагать, что они порождаются одной общей темой.

Регуляризация E-шага. Чтобы учитывать порядок слов внутри документов, удобно накладывать ограничения регуляризации на распределения $p_{tdw} = p(t\,|\,d,w)$, которые вычисляются на E-шаге для каждого слова в каждом документе. Это контекстные тематические векторы слов, которые специфицируют глобальную тематику слова $p(t\,|\,w)$ до контекста документа.

Введём регуляризатор $R(\Pi,\Phi,\Theta)$ как функцию матриц Φ, Θ и трёхмерной матрицы вспомогательных переменных $\Pi = (p_{tdw})_{T\times D\times W}$. Согласно уравнению (7), матрица Π выражается через Φ и Θ. Поэтому к регуляризатору $\tilde{R}(\Phi,\Theta) = R(\Pi(\Phi,\Theta),\Phi,\Theta)$ применима теорема 4.1. Однако систему уравнений удобно записывать через частные производные регуляризатора R, а не $\tilde{R}$.

Рассмотрим задачу максимизации регуляризованного log-правдоподобия при ограничениях неотрицательности и нормировки (5):

$$\sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{t\in T} \varphi_{wt}\theta_{td} + R(\Pi(\Phi,\Theta),\Phi,\Theta) \;\to\; \max_{\Phi,\Theta}. \tag{21}$$

Теорема 9.1 *Пусть функция $R(\Pi,\Phi,\Theta)$ непрерывно дифференцируема и не зависит от переменных p_{tdw} при $w \notin d$. Тогда точка (Φ,Θ) локального экстремума задачи (21), (5) удовлетворяет системе уравнений со вспомогательными переменными p_{tdw} и $\tilde{p}_{tdw}$, если из решения исключить нулевые столбцы матриц Φ, Θ:*

$$p_{tdw} = \underset{t\in T}{\operatorname{norm}}\big(\varphi_{wt}\theta_{td}\big);$$

$$\tilde{p}_{tdw} = p_{tdw}\left(1 + \frac{1}{n_{dw}}\left(\frac{\partial R}{\partial p_{tdw}} - \sum_{z\in T} p_{zdw}\frac{\partial R}{\partial p_{zdw}}\right)\right); \tag{22}$$

$$\varphi_{wt} = \underset{w\in W}{\operatorname{norm}}\left(\sum_{d\in D} n_{dw}\tilde{p}_{tdw} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}}\right); \tag{23}$$

$$\theta_{td} = \underset{t\in T}{\operatorname{norm}}\left(\sum_{w\in d} n_{dw}\tilde{p}_{tdw} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right). \tag{24}$$

Доказательство. Введём функцию $p_{zdw}(\Phi,\Theta) = \frac{\varphi_{wz}\theta_{zd}}{\sum_t \varphi_{wt}\theta_{td}}$ и найдём её частные производные. Для любых $t, z \in T$

$$\varphi_{wt}\frac{\partial p_{zdw}}{\partial \varphi_{wt}} = \varphi_{wt}\frac{[z{=}t]\theta_{td}\sum_u \varphi_{wu}\theta_{ud} - \theta_{td}\varphi_{wz}\theta_{zd}}{(\sum_u \varphi_{wu}\theta_{ud})^2}$$
$$= p_{tdw}[z{=}t] - p_{tdw}p_{zdw}; \tag{25}$$
$$\theta_{td}\frac{\partial p_{zdw}}{\partial \varphi_{td}} = \theta_{td}\frac{[z{=}t]\varphi_{wt}\sum_u \varphi_{wu}\theta_{ud} - \varphi_{wt}\varphi_{wz}\theta_{zd}}{(\sum_u \varphi_{wu}\theta_{ud})^2}$$
$$= p_{tdw}[z{=}t] - p_{tdw}p_{zdw}; \tag{26}$$

Заметим, что результирующие выражения (25) и (26) совпадают. Введём вспомогательную функцию Q от переменных Π, Φ, Θ:

$$Q_{tdw}(\Pi,\Phi,\Theta) = \frac{\partial R(\Pi,\Phi,\Theta)}{\partial p_{tdw}} - \sum_{z\in T} p_{zdw}\frac{\partial R(\Pi,\Phi,\Theta)}{\partial p_{zdw}}.$$

Продифференцируем суперпозицию $\tilde{R}(\Phi,\Theta) = R(\Pi(\Phi,\Theta),\Phi,\Theta)$, учитывая, что $\partial p_{zdw'}/\partial\varphi_{wt} = 0$ при $w \neq w'$; $\partial p_{zd'w}/\partial\theta_{td} = 0$ при $d \neq d'$; $\partial R/\partial p_{tdw} = 0$ при $w \notin d$:

$$\varphi_{wt}\frac{\partial \tilde{R}}{\partial \varphi_{wt}} = \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}} + \sum_{d\in D}\varphi_{wt}\sum_{z\in T}\frac{\partial R}{\partial p_{zdw}}\frac{\partial p_{zdw}}{\partial \varphi_{wt}}; \tag{27}$$

$$\theta_{td}\frac{\partial \tilde{R}}{\partial \theta_{td}} = \theta_{td}\frac{\partial R}{\partial \theta_{td}} + \sum_{w\in d}\theta_{td}\sum_{z\in T}\frac{\partial R}{\partial p_{zdw}}\frac{\partial p_{zdw}}{\partial \theta_{td}}. \tag{28}$$

Воспользовавшись (25) и (26), получим тождество

$$\varphi_{wt}\sum_{z\in T}\frac{\partial R}{\partial p_{zdw}}\frac{\partial p_{zdw}}{\partial \varphi_{wt}} = \theta_{td}\sum_{z\in T}\frac{\partial R}{\partial p_{zdw}}\frac{\partial p_{zdw}}{\partial \theta_{td}} = p_{tdw}Q_{tdw}.$$

Подставим полученные выражения в (27) и (28), которые затем подставим в систему уравнений из теоремы 4.1:

$$p_{tdw} = \underset{t\in T}{\operatorname{norm}}\big(\varphi_{wt}\theta_{td}\big);$$
$$\varphi_{wt} = \underset{w\in W}{\operatorname{norm}}\bigg(\sum_{d\in D} n_{dw}p_{tdw} + \sum_{d\in D} Q_{tdw}p_{tdw} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}}\bigg); \tag{29}$$
$$\theta_{td} = \underset{t\in T}{\operatorname{norm}}\bigg(\sum_{w\in d} n_{dw}p_{tdw} + \sum_{w\in d} Q_{tdw}p_{tdw} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\bigg). \tag{30}$$

Выделение вспомогательной переменной $\tilde{p}_{tdw}$ согласно (22) позволяет переписать уравнения (29)–(30) в требуемом виде (23)–(24).

Теорема доказана.

Таким образом, в ЕМ-алгоритме для каждого слова в документе (d,w) сначала вычисляются тематические векторы $p_{tdw} = p(t|d,w)$, затем они трансформируются в новые векторы $\tilde{p}_{tdw}$, которые подставляются в формулы М-шага (8)–(9) вместо p_{tdw}. Такой способ вычислений будем называть *регуляризацией Е-шага* или *пост-обработкой*

E-шага. Это опциональная процедура, наличие или отсутствие которой никак не влияет на реализацию остальных вычислений.

Более того, формулу пост-обработки совершенно не обязательно выводить из критерия регуляризации. Можно поступить наоборот: трансформировать последовательность тематических векторов с помощью эвристической пост-обработки, например, разреживания, сглаживания или сегментирования. Фактически, это будет соответствовать регуляризации при некотором критерии $R(\Pi)$, который можно даже не выписывать в явном виде.

Данный подход использовался в [47] для повышения качества тематической сегментации документов.

Линейная тематизация документов. Вычисление тематического вектора документа $\theta_d = (\theta_{td})_{t\in T}$ в EM-алгоритме требует многих итераций по документу. В [25] предлагается вычислять θ_d за один линейный проход по документу с помощью явной формулы $\theta_{td}(\Phi)$. Фактически, это ограничение-равенство играет роль регуляризатора. Оно может быть получено из формулы M-шага или из формулы полной вероятности, где распределение $p(t)$ полагается фиксированным:

$$\theta_{td}(\Phi) = \sum_{w\in d} p(t|w)\, p(w|d) = \sum_{w\in d} \frac{n_{dw}}{n_d} \operatorname*{norm}_{t\in T}\big(\varphi_{wt} p(t)\big).$$

Теорема 9.2 *Пусть функции $\theta_{td}(\Phi)$ и $R(\Phi,\Theta)$ непрерывно дифференцируемы. Тогда точка Φ локального экстремума задачи (6), (5) с ограничениями-равенствами $\theta_{td} = \theta_{td}(\Phi)$ удовлетворяет системе уравнений со вспомогательными переменными $p_{tdw} = p(t\,|\,d,w)$, n_{td} и p'_{tdw}, если из решения исключить нулевые столбцы матриц Φ, Θ:*

$$\begin{aligned}
p_{tdw} &= \operatorname*{norm}_{t\in T}\big(\varphi_{wt}\theta_{td}\big);\\
n_{td} &= \sum_{w\in d} n_{dw} p_{tdw} + \theta_{td}\frac{\partial R}{\partial \theta_{td}};\\
p'_{tdw} &= p_{tdw} + \frac{\varphi_{wt}}{n_{dw}} \sum_{s\in T} \frac{n_{sd}}{\theta_{sd}} \frac{\partial \theta_{sd}}{\partial \varphi_{wt}};\\
\varphi_{wt} &= \operatorname*{norm}_{w\in W}\Big(\sum_{d\in D} n_{dw} p'_{tdw} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}}\Big).
\end{aligned}$$

Таким образом, модификация EM-алгоритма, как и в случае пост-обработки E-шага, сводится к трасформации тематических векторов p_{tdw} в p'_{tdw}, которые подставляются в обычную формулу M-шага для матрицы Φ, не влияя на реализацию остальных вычислений.

Эксперименты [25] на трёх свободно доступных текстовых коллекциях показали, что линейная тематизация документов не только вычисляется быстрее, но и играет роль регуляризатора, улучшая качество модели по пяти характеристикам разреженности, различности, информативности и когерентности тем.

Линейная тематизация открывает возможности для быстрого вычисления тематических векторов контекстов слов и перехода от гипотезы мешка слов к обработке текста как последовательности.

10 Обсуждения и выводы

Сотни байесовских тематических моделей, описанных в тысячах статей за последние два десятилетия, могут быть дебаесизированы и выведены прямолинейно, буквально в одну строку, из теоремы о максимизации гладкой функции на единичных симплексах. Может возникнуть вопрос, как так получилось, что эта возможность не была замечена сразу. Ведь байесовский вывод, трудоёмкий и уникальный для каждой модели, приносит исследователям много технических неудобств.

Многие области анализа данных и машинного обучения, включая обработку изображений и сигналов, развивались по общему сценарию. Сначала формализация модели и оптимизационной задачи; затем её дополнение различными структурами, вспомогательными критериями и регуляризаторами; и только в последнюю очередь переход к байесовской регуляризации. Этот переход нужен тогда, когда возникает практическая потребность оценивания не только самих параметров модели, но и их апостериорных распределений.

В тематическом моделировании типичный сценарий развития был нарушен, и сообщество перешло к методам байесовского обучения минуя этап развития в рамках классической регуляризации. Это тем более парадоксально, что в практике тематического моделирования апостериорные распределения используются только для получения точечных оценок максимального правдоподобия.

Аддитивная регуляризация есть попытка восполнить этот пробел. Возможно, попытка запоздалая, поскольку фокус интереса сообщества уже переключился на глубокие нейросетевые модели языка, модели внимания и архитектуры трансформеров. Тематическое моделирование теперь больше сосредоточено на интеграции с нейронными сетями в поисках возможностей для «объединения лучшего от двух миров» [61].

Оба вида моделей, нейросетевые и тематические, строят векторные представления слов и текстов.

Оба вида моделей тяготеют к гомогенизации [39], то есть использованию единого векторного пространства для описания разнородных объектов по данным об их взаимодействиях. Выше мы показали, как это реализуется в гиперграфовых тематических моделях.

Оба вида моделей генерируют как глобальные эмбединги, так и локальные для каждого слова в его контексте. Выше мы показали, как строятся тематические модели последовательного текста. Нейросетевые модели намного сложнее, их эмбединги способны вобрать в себя всю полноту информации о связях между словами, причём мы даже не понимаем, какие именно связи и как именно учтены. Тематические модели намного проще, их эмбединги учитывают только лексическую сочетаемость слов, но сохраняют интерпретируемость. Свойство покоординатной интерпретируемости является прямым следствием того, что тематические эмбединги являются неотрицательными нормированными векторами.

Отказ от байесовского вывода тематических моделей сближает их с нейросетевыми моделями, делает возможным их более глубокую интеграцию. Любой векторный параметр нейронной сети, если на

него наложить ограничения неотрицательности и нормировки, можно обучать с помощью формулы мультипликативного градиентного шага из теоремы о максимизации функции на единичных симплексах. Это перспективная возможность для будущих исследований.

Работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований (проект 20-07-00936).

Список литературы

[1] *Apishev M., Koltcov S., Koltsova O., Nikolenko S., Vorontsov K.* Additive regularization for topic modeling in sociological studies of user-generated text content // MICAI 2016, 15th Mexican International Conference on Artificial Intelligence. — Vol. 10061. — Springer, Lecture Notes in Artificial Intelligence, 2016. — Pp. 166–181.

[2] *Apishev M., Koltcov S., Koltsova O., Nikolenko S., Vorontsov K.* Mining ethnic content online with additively regularized topic models // *Computacion y Sistemas.* — 2016. — Vol. 20, no. 3. — Pp. 387–403.

[3] *Apishev M. A., Vorontsov K. V.* Learning topic models with arbitrary loss // Proceeding of the 26th Conference of FRUCT (Finnish-Russian University Cooperation in Telecommunications) Association. — 2020. — Pp. 30–37.

[4] Attention is all you need / A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, I. Polosukhin // Advances in Neural Information Processing Systems 30 / Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett. — Curran Associates, Inc., 2017. — Pp. 5998–6008.

[5] *Balikas G., Amini M., Clausel M.* On a topic model for sentences // Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval. — SIGIR '16. — New York, NY, USA: ACM, 2016. — Pp. 921–924.

[6] *Belyy A. V., Seleznova M. S., Sholokhov A. K., Vorontsov K. V.* Quality evaluation and improvement for hierarchical topic modeling // Computational Linguistics and Intellectual Technologies. Dialogue 2018. — 2018. — Pp. 110–123.

[7] *Blei D. M.* Probabilistic topic models // *Communications of the ACM.* — 2012. — Vol. 55, no. 4. — Pp. 77–84.

[8] *Blei D. M., Ng A. Y., Jordan M. I.* Latent Dirichlet allocation // *Journal of Machine Learning Research.* — 2003. — Vol. 3. — Pp. 993–1022.

[9] *Bodrunova S., Koltsov S., Koltsova O., Nikolenko S. I., Shimorina A.* Interval semi-supervised LDA: Classifying needles in a haystack // MICAI (1) / Ed. by F. C. Espinoza, A. F. Gelbukh, M. Gonzalez-Mendoza. — Vol. 8265 of *Lecture Notes in Computer Science.* — Springer, 2013. — Pp. 265–274.

[10] *Boyd-Graber J., Hu Y., Mimno D.* Applications of topic models // *Foundations and Trends® in Information Retrieval.* — 2017. — Vol. 11, no. 2-3. — Pp. 143–296.

[11] *Bulatov V., Egorov E., Veselova E., Polyudova D., Alekseev V., Goncharov A., Vorontsov K.* TopicNet: Making additive regularisation for topic modelling accessible // Proceedings of The 12th Conference on Language Resources and Evaluation (LREC 2020). — 2020. — Pp. 6745–6752.

[12] *Chen B.* Word topic models for spoken document retrieval and transcription. — 2009. — Vol. 8, no. 1. — Pp. 2:1–2:27.

[13] *Chen R., Hua Q., Chang Y.-S., Wang B., Zhang L., Kong X.* A survey of collaborative filtering-based recommender systems: From traditional methods to hybrid methods based on social networks // *IEEE Access.* — 2018. — Vol. 6. — Pp. 64301–64320.

[14] *Chirkova N. A., Vorontsov K. V.* Additive regularization for hierarchical multimodal topic modeling // *Journal Machine Learning and Data Analysis.* — 2016. — Vol. 2, no. 2. — Pp. 187–200.

[15] *Churchill R., Singh L.* The evolution of topic modeling // *ACM Comput. Surv.* — 2022. — Vol. 54, no. 10s. — 35 pp.

[16] *Devlin J., Chang M.-W., Lee K., Toutanova K.* BERT: Pre-training of deep bidirectional transformers for language understanding // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). — Minneapolis, Minnesota: Association for Computational Linguistics, 2019. — Pp. 4171–4186.

[17] *Dudarenko M. A.* Regularization of multilingual topic models // *Vychisl. Metody Programm. (Numerical methods and programming).* — 2015. — Vol. 16. — Pp. 26–38.

[18] *El-Kishky A., Song Y., Wang C., Voss C. R., Han J.* Scalable topical phrase mining from text corpora // *Proc. VLDB Endowment.* — 2014. — Vol. 8, no. 3. — Pp. 305–316.

[19] *Feldman D. G., Sadekova T. R., Vorontsov K. V.* Combining facts, semantic roles and sentiment lexicon in a generative model for opinion mining // Computational Linguistics and Intellectual Technologies. Dialogue 2020. — 2020. — Pp. 268–283.

[20] *Frei O., Apishev M.* Parallel non-blocking deterministic algorithm for online topic modeling // AIST'2016, Analysis of Images, Social networks and Texts. — Vol. 661. — Springer International Publishing Switzerland, Communications in Computer and Information Science (CCIS), 2016. — Pp. 132–144.

[21] *Harris Z.* Distributional structure // *Word.* — 1954. — Vol. 10, no. 23. — Pp. 146–162.

[22] *Hofmann T.* Probabilistic latent semantic indexing // Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. — New York, NY, USA: ACM, 1999. — Pp. 50–57.

[23] *Ianina A., Vorontsov K.* Regularized multimodal hierarchical topic model for document-by-document exploratory search // Proceeding Of The 25th Conference Of FRUCT (Finnish-Russian University Cooperation in Telecommunications) Association. The seminar on Intelligence, Social Media and Web (ISMW). Helsinki, Finland, November 5–8, 2019. / Ed. by S. Balandin, V. Niemi, T. Tutina. — 2019. — Pp. 131–138.

[24] *Ianina A. O., Vorontsov K. V.* Hierarchical interpretable topical embeddings for exploratory search and real-time document tracking // *International Journal of Embedded and Real-Time Communication Systems (IJERTCS).* — 2020. — Vol. 11, no. 4. — 19 pp.

[25] *Irkhin I. A., Bulatov V. G., Vorontsov K. V.* Additive regularization of topic models with fast text vectorization // *Computer Research and Modeling.* — 2020. — Vol. 12, no. 6. — Pp. 1515–1528.

[26] *Irkhin I. A., Vorontsov K. V.* Convergence of the algorithm of additive regularization of topic models // *Trudy Instituta Matematiki i Mekhaniki UrO RAN.* — 2020. — Vol. 26, no. 3. — Pp. 56–68.

[27] *Jameel S., Lam W.* An N-gram topic model for time-stamped documents // 35th European Conference on Information Retrieval, ECIR-2013, Moscow, Russia, 24-27 March 2013. — Lecture Notes in Computer Science (LNCS), Springer Verlag-Germany, 2013. — Pp. 292–304.

[28] *Jelodar H., Wang Y., Yuan C., Feng X., Jiang X., Li Y., Zhao L.* Latent dirichlet allocation (LDA) and topic modeling: models, applications, a survey // *Multimedia Tools and Applications.* — 2019. — Vol. 78, no. 11. — Pp. 15169–15211.

[29] *Khodorchenko M., Butakov N., Sokhin T., Teryoshkin S.* Surrogate-based optimization of learning strategies for additively regularized topic models // *Logic Journal of the IGPL.* — 2022. — jzac019.

[30] *Khodorchenko M., Teryoshkin S., Sokhin T., Butakov N.* Optimization of learning strategies for artm-based topic models // Hybrid Artificial Intelligent Systems / Ed. by E. A. de la Cal, J. R. Villar Flecha, H. Quintián, E. Corchado. — Springer International Publishing, 2020. — Pp. 284–296.

[31] *Kochedykov D. A., Apishev M. A., Golitsyn L. V., Vorontsov K. V.* Fast and modular regularized topic modelling // Proceeding of the 21st Conference Of FRUCT (Finnish-Russian University Cooperation in Telecommunications) Association. The seminar on Intelligence, Social Media and Web (ISMW). Helsinki, Finland, November 6–10, 2017. — IEEE, 2017. — Pp. 182–193.

[32] *Koltcov S., Koltsova O., Nikolenko S.* Latent Dirichlet allocation: Stability and applications to studies of user-generated content // Proceedings of the 2014 ACM Conference on Web Science. — WebSci'14. — New York, NY, USA: ACM, 2014. — Pp. 161–165.

[33] Language models are few-shot learners // Advances in Neural Information Processing Systems / Ed. by H. Larochelle, M. Ranzato,

R. Hadsell, M. Balcan, H. Lin. — Vol. 33. — Curran Associates, Inc., 2020. — Pp. 1877–1901.

[34] *M. A. Basher A. R., Fung B. C. M.* Analyzing topics and authors in chat logs for crime investigation // *Knowledge and Information Systems.* — 2014. — Vol. 39, no. 2. — Pp. 351–381.

[35] *Mei Q., Shen X., Zhai C.* Automatic labeling of multinomial topic models // Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. — New York, NY, USA: Association for Computing Machinery, 2007. — Pp. 490–499.

[36] *Mikolov T., Chen K., Corrado G., Dean J.* Efficient estimation of word representations in vector space // *CoRR.* — 2013. — Vol. abs/1301.3781.

[37] *Mikolov T., Sutskever I., Chen K., Corrado G., Dean J.* Distributed representations of words and phrases and their compositionality // *CoRR.* — 2013. — Vol. abs/1310.4546.

[38] *Nikolenko S. I., Koltcov S., Koltsova O.* Topic modelling for qualitative studies // *Journal of Information Science.* — 2017. — Vol. 43, no. 1. — Pp. 88–102.

[39] On the opportunities and risks of foundation models / R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein et al. // *CoRR.* — 2021. — Vol. abs/2108.07258.

[40] *Paul M. J., Dredze M.* Drug extraction from the web: Summarizing drug experiences with multi-dimensional topic models // Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9–14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA. — 2013. — Pp. 168–178.

[41] *Paul M. J., Dredze M.* Discovering health topics in social media using topic models // *PLoS ONE.* — 2014. — Vol. 9, no. 8.

[42] *Popov A., Bulatov V., Polyudova D., Veselova E.* Unsupervised dialogue intent detection via hierarchical topic model // Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019). — Varna, Bulgaria: INCOMA Ltd., 2019. — Pp. 932–938.

[43] *Potapenko A., Popov A., Vorontsov K.* Interpretable probabilistic embeddings: bridging the gap between topic models and neural networks // Communications in Computer and Information Science, vol 789. AINL-6: Artificial Intelligence and Natural Language Conference, St. Petersburg, Russia, September 20-23, 2017. — Springer, Cham, 2017. — Pp. 167–180.

[44] *Rosen-Zvi M., Griffiths T., Steyvers M., Smyth P.* The author-topic model for authors and documents // Proceedings of the 20th conference on Uncertainty in artificial intelligence. — UAI '04. — Arlington, Virginia, United States: AUAI Press, 2004. — Pp. 487–494.

[45] *Rubin T. N., Chambers A., Smyth P., Steyvers M.* Statistical topic models for multi-label document classification // *Machine Learning.* — 2012. — Vol. 88, no. 1-2. — Pp. 157–208.

[46] *Sharma A., Pawar D. M.* Survey paper on topic modeling techniques to gain usefull forcasting information on violant extremist activities over cyber space // *International Journal of Advanced Research in Computer Science and Software Engineering.* — 2015. — Vol. 5, no. 12. — Pp. 429–436.

[47] *Skachkov N. A., Vorontsov K. V.* Improving topic models with segmental structure of texts // Computational Linguistics and Intellectual Technologies. Dialogue 2018. — 2018. — Pp. 652–661.

[48] *Tan Y., Ou Z.* Topic-weak-correlated latent Dirichlet allocation // 7th International Symposium Chinese Spoken Language Processing (ISCSLP). — 2010. — Pp. 224–228.

[49] *Tikhonov A. N., Arsenin V. Y.* Solution of ill-posed problems. — W. H. Winston, Washington, DC, 1977.

[50] *Vinokourov A., Girolami M.* A probabilistic hierarchical clustering method for organising collections of text documents // Proceedings 15th International Conference on Pattern Recognition. ICPR-2000. — Vol. 2. — 2000. — Pp. 182–185 vol.2.

[51] *Vorontsov K., Frei O., Apishev M., Romov P., Suvorova M., Yanina A.* Non-bayesian additive regularization for multimodal topic modeling of large collections // Proceedings of the 2015 Workshop on Topic Models: Post-Processing and Applications. — New York, NY, USA: ACM, 2015. — Pp. 29–37.

[52] *Vorontsov K. V.* Additive regularization for topic models of text collections // *Doklady Mathematics.* — 2014. — Vol. 89, no. 3. — Pp. 301–304.

[53] *Vorontsov K. V., Frei O. I., Apishev M. A., Romov P. A., Suvorova M. A.* BigARTM: Open source library for regularized multimodal topic modeling of large collections // AIST'2015, Analysis of Images, Social networks and Texts. — Springer International Publishing Switzerland, Communications in Computer and Information Science (CCIS), 2015. — Pp. 370–384.

[54] *Vorontsov K. V., Potapenko A. A.* Tutorial on probabilistic topic modeling: Additive regularization for stochastic matrix factorization // AIST'2014, Analysis of Images, Social networks and Texts. — Vol. 436. — Springer International Publishing Switzerland, Communications in Computer and Information Science (CCIS), 2014. — Pp. 29–46.

[55] *Vorontsov K. V., Potapenko A. A.* Additive regularization of topic models // *Machine Learning, Special Issue on Data Analysis and Intelligent Optimization with Applications.* — 2015. — Vol. 101, no. 1. — Pp. 303–323.

[56] *Vulic I., De Smet W., Tang J., Moens M.-F.* Probabilistic topic modeling in multilingual settings: an overview of its methodology and

applications // *Information Processing & Management.* — 2015. — Vol. 51, no. 1. — Pp. 111–147.

[57] *Wallach H. M.* Topic modeling: Beyond bag-of-words // Proceedings of the 23rd International Conference on Machine Learning. — ICML '06. — New York, NY, USA: ACM, 2006. — Pp. 977–984.

[58] *Wang X., McCallum A., Wei X.* Topical n-grams: Phrase and topic discovery, with an application to information retrieval // Proceedings of the 2007 Seventh IEEE International Conference on Data Mining. — Washington, DC, USA: IEEE Computer Society, 2007. — Pp. 697–702.

[59] *Yanina A., Golitsyn L., Vorontsov K.* Multi-objective topic modeling for exploratory search in tech news // Communications in Computer and Information Science, vol 789. AINL-6: Artificial Intelligence and Natural Language Conference, St. Petersburg, Russia, September 20-23, 2017 / Ed. by A. Filchenkov, L. Pivovarova, J. Žižka. — Springer International Publishing, Cham, 2018. — Pp. 181–193.

[60] *Zavitsanos E., Paliouras G., Vouros G. A.* Non-parametric estimation of topic hierarchies from texts with hierarchical Dirichlet processes // *Journal of Machine Learning Research.* — 2011. — Vol. 12. — Pp. 2749–2775.

[61] *Zhao H., Phung D., Huynh V., Jin Y., Du L., Buntine W.* Topic modelling meets deep neural networks: A survey // Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21 / Ed. by Z.-H. Zhou. — International Joint Conferences on Artificial Intelligence Organization, 8 2021. — Pp. 4713–4720.

[62] *Zhao W. X., Jiang J., Weng J., He J., Lim E.-P., Yan H., Li X.* Comparing Twitter and traditional media using topic models // Proceedings of the 33rd European Conference on Advances in Information Retrieval. — ECIR'11. — Berlin, Heidelberg: Springer-Verlag, 2011. — Pp. 338–349.

[63] *Zuo Y., Zhao J., Xu K.* Word network topic model: A simple but general solution for short and imbalanced texts // *Knowledge and Information Systems.* — 2016. — Vol. 48, no. 2. — Pp. 379–398.