Вероятностное тематическое моделирование: теория регуляризации ARTM и библиотека с открытым кодом `BigARTM`

Воронцов Константин Вячеславович

vokov@forecsys.ru

Московский государственный университет им. М. В. Ломоносова,

Московский физико-технический институт (государственный университет),

Федеральный исследовательский центр «Информатика и управление» РАН

15 ноября 2024 г.

Аннотация

Вероятностное тематическое моделирование — это технология автоматической обработки текстов, активно развивающаяся последние два десятилетия. Тематические модели используются для разведочного анализа больших текстовых коллекций, информационного поиска и решения разнообразных задач текстовой аналитики. Книга охватывает основные типы тематических моделей. Математический аппарат классической не-байесовской регуляризации позволяет существенно упростить изложение по сравнению с байесовским обучением, обычно используемым в научной литературе. Большое внимание уделяется формализации постановок задач на языке оптимизационных критериев. Математическая теория аддитивной регуляризации (ARTM) приводит к модульной технологии моделирования, которая реализована в библиотеке с открытым кодом `BigARTM`. Отдельная глава посвящена работе с этой библиотекой.

Книга адресована специалистам в области машинного обучения и автоматической обработки текстов, преподавателям, студентам, аспирантам, исследователям и инженерам-практикам. Предполагается, что читатель владеет основами теории вероятностей, линейной алгебры, математической статистики, численных методов оптимизации, языка программирования Python в объёме курсов бакалавриата технических университетов.

Содержание

1 Введение

Тематическое моделирование — это одна из современных технологий *обработки естественного языка* (natural language processing, NLP), активно развивающаяся с конца 90-х годов. Тематическая модель коллекции текстовых документов определяет, к каким темам относится каждый документ, и какие слова образуют каждую тему. Тематическое моделирование не претендует на полноценное *понимание естественного языка* (natural language understanding, NLU), однако выявление тематики можно считать определённым шагом в этом направлении.

Тематическое моделирование принято относить к *машинному обучению без учителя* (unsupervised machine learning), поскольку темы строятся автоматически по текстовым данным. Для этого не требуется ни разметки, ни тезаурусов, ни баз экспертных знаний. Существуют продвинутые тематические модели, способные учитывать такого рода данные для улучшения тем и решения трудных задач текстовой аналитики. Такие модели тоже рассматриваются в данной книге.

Тематическое моделирование похоже на *кластеризацию документов*. Отличие в том, что при обычной «жёсткой» кластеризации (hard clustering) документ целиком относится к одному кластеру, тогда как тематическая модель осуществляет *мягкую кластеризацию* (soft clustering), распределяя содержимое документа по нескольким кластерам-темам. Тематическое моделирование называют также *мягкой би-кластеризацией*, поскольку каждое слово также распределяется по темам.

Вероятностная тематическая модель (probabilistic topic model, PTM) определяет вероятности тем в каждом документе и вероятности слов в каждой теме. Такие модели предсказывают вероятности появления слов в документах, но делают это не настолько хорошо, как глубокие нейронные сети типа BERT [58] или GPT [98]. Зато они намного проще и обладают свойством интерпретируемости.

Тематическая модель, как и нейросетевая, преобразует слова и тексты в их *векторные представления* или эмбединги (embedding). Нейросетевые эмбединги не интерпретируемы, мы не понимаем смысла координат в этих векторах. Тематический эмбединг — это вектор вероятностей тем. В каждой теме есть наиболее частотные слова, и если модель построена хорошо, то они оказываются связанными друг с другом по смыслу. Глядя на них, можно сказать, о чём эта тема, составить её текстовое описание, дать ей название [118]. Наиболее ценное свойство тематических моделей в том, что коллекция сама собой кластеризуется на интерпретируемые темы.

Мы почти не будем касаться вопроса, как «объединить лучшее от двух миров» — предсказательную и генеративную способность нейросетевых моделей языка и интерпретируемость тематических моделей. Это открытая научная проблема на момент написания данного предисловия. Хотелось бы верить, что математический инструментарий, предлагаемый в данной книге, поможет мотивированным читателям в решении этой новой увлекательной задачи.

Вероятностные тематические модели находят множество применений. Это выявление трендов в новостных потоках, патентных базах, научных публикациях [208, 174], многоязычный информационный поиск [184, 183], классификация и категоризация документов [156, 212], тематическая сегментация текстов [195, 153], суммаризация текстов [106], поиск тематических сообществ в социальных сетях [211, 176, 145, 42], тегирование веб-страниц [95], обнаружение текстового спама [16]. Существуют и не-текстовые приложения тематического моделирования в анали-

зе изображений и видеопотоков [78, 105, 67, 175], в рекомендательных системах [202, 190, 100, 205, 204], в популяционной генетике [148], в биоинформатике для анализа нуклеотидных [96] и аминокислотных последовательностей [163, 94]. Другие приложения тематических моделей упоминаются в обзорах [55, 36, 44, 152, 86, 51].

Построение тематической модели является некорректно поставленной оптимизационной задачей, имеющей бесконечно много решений. Согласно теории регуляризации А. Н. Тихонова [17], решение такой задачи возможно доопределить и сделать устойчивым. Для этого к оптимизационному критерию добавляется *регуляризатор* — дополнительный критерий, учитывающий специфические особенности данных или предметной области. Тематические модели обладают огромным «запасом решений». Можно добавить регуляризатор, сильно улучшив один критерий качества, затем добавить второй и улучшить модель по другому критерию, и так несколько раз.

Аддитивная регуляризация тематических моделей (additive regularization of topic models, ARTM) — это многокритериальный подход, в котором модель оптимизируется по сумме критериев [5, 181]. ARTM позволяет строить модели с требуемыми свойствами, объединяя регуляризаторы от различных моделей. Для обучения любых моделей и их комбинаций используется один и тот же алгоритм, к которому регуляризаторы подключаются как модули [179, 68, 92]. Модульная технология реализована в библиотеке `BigARTM` с открытым кодом, `http://bigartm.org` [179, 68].

Подчеркнём, что ARTM не является ещё одной моделью или методом — это общий подход к построению и комбинированию любых тематических моделей.

До настоящего времени доминирующим подходом в тематическом моделировании оставалось байесовское обучение. В отличие от ARTM, в нём не получается выделять унифицируемые модули-регуляризаторы. Для каждой байесовской модели приходится заново выводить вычислительные формулы, реализовывать и тестировать программный код. Из-за математической сложности байесовского вывода в статьях часто опускаются важные для понимания детали. Эти барьеры препятствуют широкому использованию тематического моделирования. Как следствие, в индустрии анализа текстов редко применяются модели сложнее морально устаревшей модели LDA (Latent Dirichlet Allocation) [40]. Сотни перспективных моделей так и остаются «академическими исследованиями для одной статьи».

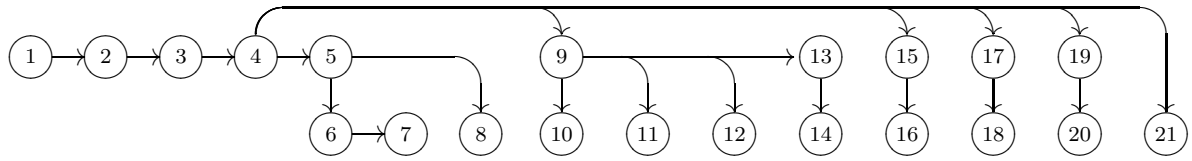
Основная цель книги — показать огромное разнообразие тематических моделей, сосредоточившись на постановках оптимизационных задач. Показать, что классическая (не-байесовская) регуляризация является не менее выразительным средством моделирования, чем байесовское обучение. На этом языке возможно не только строить и комбинировать тематические модели, но также объяснять их намного лаконичнее, без «заметания под ковёр» сложной математики. Сопоставимый по охвату и обстоятельности обзор байесовских моделей занял бы много сотен страниц.

Основоположник современной теории вероятностей А. Н. Колмогоров не раз отмечал в своих лекциях: «представляется важной задача освобождения всюду, где это возможно, от излишних вероятностных допущений» [13, стр. 252]. Оказалось, что освобождение от избыточной байесовской постановки приводит к более изящной теории тематического моделирования. Она основана на одной лемме, которую мы докажем в главе 3. Все алгоритмы тематического моделирования выводятся из неё единообразно, практически в одно действие. Это должны по достоинству оценить исследователи, подбиравшие уникальные техники байесовского вывода для тематических моделей, хоть немного более сложных, чем LDA.

Главы 2–5 являются базовыми. В главах 6–7 излагаются основы байесовской теории; она не обязательна для понимания последующего материала, но может быть полезна при чтении научных статей. В главах 8–16 в терминах регуляризации описываются различные виды тематических моделей. Они слабо связаны друг с другом, их можно читать в произвольном порядке или использовать как путеводитель по литературе. Глава 9 вводит понятие модальности, существенно расширяющее спектр приложений. В главе 10 предлагается широкое обобщение тематических моделей для обработки транзакционных данных произвольной природы. Глава 13 посвящена иерархическим тематическим моделям, в которых темы делятся на подтемы. Главы 17–18 описывают критерии качества тематических моделей. В главе 19 рассматриваются особенности реализации алгоритмов, в том числе приёмы ускорения сходимости и распараллеливания. Глава 20 содержит начальные сведения о библиотеке `BigARTM`. В главе 21 обсуждается применение тематического моделирования для разведочного информационного поиска и социо-гуманитарных исследований. В главе 22 подводятся итоги, развенчиваются мифы и обсуждаются открытые проблемы.

В начале каждой главы приводится краткая мотивация и неформальное пояснение, что будет происходить далее, и почему это важно. В конце главы тезисно формулируются наиболее важные выводы.

На схеме зависимости глав в верхнем ряду расположены основные главы, в нижнем — дополнительные. При построении учебного курса дополнительные главы можно брать выборочно или оставлять для самостоятельного изучения.



Книга написана по материалам курса, читаемого студентам факультета ВМК МГУ с 2013 года и факультета ФУПМ МФТИ с 2019 года. Презентации и другие материалы курса можно найти на странице вики-ресурса `www.MachineLearning.ru` «Вероятностные тематические модели (курс лекций, К.В.Воронцов)».

Константин Воронцов
26 июня 2023 г.
15 ноября 2024 г.

2 Основы тематического моделирования

В этой главе мы введём основные понятия и поставим задачу тематического моделирования как задачу приближённого низкорангового стохастического матричного разложения. Обсудим свойства этой задачи и сформулируем цели тематического моделирования. С помощью элементарных инструментов — формулы Байеса и частотных оценок условных вероятностей — получим итерационный процесс, называемый EM-алгоритмом. Почему он действительно решает поставленную задачу, узнаем в двух следующих главах.

Предварительная обработка текста. Перед построением тематических моделей текст естественного языка обычно подвергается серии преобразований.

Лемматизация — это приведение каждого слова в документе к его нормальной форме. В русском языке нормальными формами считаются: для существительных — именительный падеж, единственное число; для прилагательных — именительный падеж, единственное число, мужской род; для глаголов, причастий, деепричастий — глагол в инфинитиве. Хорошими лемматизаторами для русского языка считаются последние версии `mystem` и `pymorphy`.

Стемминг — это отбрасывание окончаний и других изменяемых частей слов. Он подходит для английского языка, для русского предпочтительна лемматизация.

Удаление стоп-слов. Это частые слова, встречающиеся в текстах любой тематики — предлоги, союзы, числительные, местоимения, некоторые глаголы, прилагательные, наречия. Число таких слов обычно варьируется в пределах нескольких сотен. Они бесполезны для тематических моделей. Их отбрасывание почти не влияет на объём словаря, но может приводить к заметному сокращению длины текстов.

Удаление редких слов и строк, не являющихся словами естественного языка (например, содержащих цифры или спецсимволы), помогает во много раз сокращать объём словаря, снижая затраты времени и памяти на построение моделей. Редкие слова, как правило, не влияют на тематику коллекции.

Выделение ключевых фраз — характерных словосочетаний и терминов предметной области — используется для улучшения интерпретируемости тем. Выделять их можно с помощью тезаурусов [14] или методов автоматического выделения терминов (automatic term extraction, ATE), не требующих привлечения экспертов [64, 107, 160].

Распознавание именованных сущностей (named entities recognition, NER). Это названия объектов реального мира, относящихся к определённым категориям: персоны, организации, геолокации, события, даты, и т. д. В каждой предметной области могут быть свои категории: болезни, методы лечения, химические вещества, виды растений и животных, небесные тела, изделия, товары, и т. д. Для распознавания именованных сущностей используются различные методы машинного обучения [129, 97, 135].

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

Гипотеза о существовании тем. Каждое вхождение терма w в документ d связано с некоторой темой t из заданного конечного множества T. Коллекция документов представляет собой последовательность троек $\Omega_n = \{(w_i, d_i, t_i) \mid i = 1, \ldots, n\}$. Термы w_i и документы d_i являются наблюдаемыми переменными, темы t_i не известны и являются *латентными* (скрытыми) переменными.

Гипотеза «мешка слов». Порядок термов в документах не важен для выявления тематики, то есть тематику документа можно узнать даже после произвольной перестановки термов, хотя для человека такой текст потеряет смысл. Это предположение называют гипотезой «мешка слов» (bag of words). Порядок документов в коллекции также не имеет значения — это предположение называют гипотезой «мешка документов». Гипотеза «мешка слов» позволяет перейти к компактному представлению документа как *мультимножества* — подмножества термов $d \subset W$, в котором каждый терм $w \in d$ повторён n_{dw} раз.

Гипотеза о вероятностном порождении данных. Множество $\Omega = D \times W \times T$ является конечным *вероятностным пространством* с неизвестной функцией вероятности $p(d, w, t)$. Коллекция документов является выборкой троек (d_i, w_i, t_i), порождаемых случайно и независимо друг от друга из распределения $p(d, w, t)$. Это предположение является вероятностным уточнением гипотезы «мешка слов».

Благодаря предположению о независимости, реализовавшуюся выборку Ω_n элементов из Ω можно рассматривать как новое вероятностное пространство с n равновероятными элементарными исходами. В пространстве Ω_n легко находить вероятности различных событий, причём они совпадают с частотными оценками вероятностей тех же событий в пространстве Ω. В частности, в пространстве Ω_n выражение

$$\hat{p}(d, w, t) = \frac{1}{n} \sum_{i=1}^{n} [d_i = d]\,[w_i = w]\,[t_i = t]$$

равно вероятности события «терм w документа d связан с темой t», а в пространстве Ω оно равно выборочной частотной оценке вероятности того же события.

Договоримся в дальнейшем записывать все вероятности в пространстве Ω, если не оговорено иного. Многие выкладки будут справедливы в обоих пространствах. Пространство Ω_n имеет формальное ограничение — оно строится по фиксированной коллекции. Если в коллекцию добавляются новые документы, то пространство Ω_n изменяется, тогда как пространство Ω можно полагать неизменным.

Гипотеза условной независимости. Появление термов в документе d по теме t зависит от темы, но не зависит от документа d, и описывается общим для всех документов распределением $p(w \mid t)$:

$$p(w \mid d, t) = p(w \mid t). \tag{1}$$

Вероятностная тематическая модель. Согласно формуле полной вероятности и гипотезе условной независимости, распределение термов в документе $p(w \mid d)$ описывается *вероятностной смесью* распределений термов в темах $\varphi_{wt} = p(w \mid t)$, взятых

Алгоритм 1. Вероятностный процесс порождения коллекции документов.

Вход: распределения $p(w\,|\,t)$, $p(t\,|\,d)$; длины документов n_d;
Выход: выборка пар (d_i, w_i), $i = 1, \ldots, n$;

$i := 0$;
для всех $d \in D$
  для всех $j = 1, \ldots, n_d$
    $i := i + 1$; $d_i := d$;
    выбрать случайную тему t_i из распределения $p(t\,|\,d_i)$;
    выбрать случайный терм w_i из распределения $p(w\,|\,t_i)$;

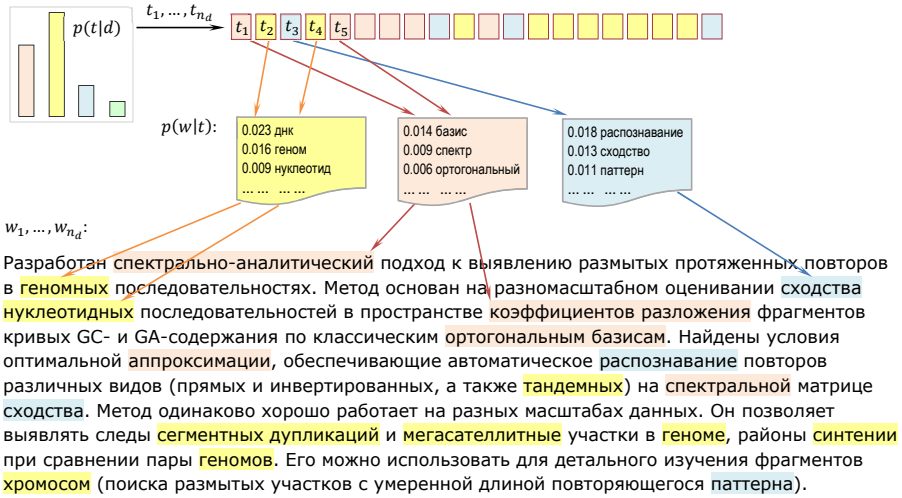


Рис. 1: Процесс порождения текстовой коллекции вероятностной тематической моделью (2): в каждой позиции i документа d_i сначала порождается тема $t_i \sim p(t\,|\,d_i)$, затем терм $w_i \sim p(w\,|\,t_i)$.

с весами, равными вероятностям тем в документах $\theta_{td} = p(t\,|\,d)$:

$$p(w\,|\,d) = \sum_{t\in T} p(w\,|\,t,d)\,p(t\,|\,d) = \sum_{t\in T} p(w\,|\,t)\,p(t\,|\,d) = \sum_{t\in T} \varphi_{wt}\theta_{td}. \qquad (2)$$

Вероятностная модель (2) описывает процесс порождения коллекции по известным распределениям $p(w\,|\,t)$ и $p(t\,|\,d)$. Этот процесс показан в алгоритме 1 и на рис. 1.

Задача тематического моделирования — это обратная задача: по заданной коллекции D требуется найти параметры φ_{wt} и θ_{td}, при которых тематическая модель (2) хорошо приближает частотные оценки условных вероятностей $\hat{p}(w\,|\,d) = p_{dw} = \frac{n_{dw}}{n_d}$.

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

Во-первых, это задача приближённого низкорангового стохастического матричного разложения. Чтобы пояснить все эти понятия, перепишем равенство (2) в матричном виде $P \approx \Phi\Theta$, где $P = (p_{dw})_{W\times D}$ — *матрица частот термов в документах*, $\Phi = (\varphi_{wt})_{W\times T}$ — *матрица термов тем* и $\Theta = (\theta_{td})_{T\times D}$ — *матрица тем документов*, см. рис. 2. Матрица P известна, это исходные данные. Правая часть равенства представляет собой произведение двух неизвестных матриц. Во всех трёх матрицах

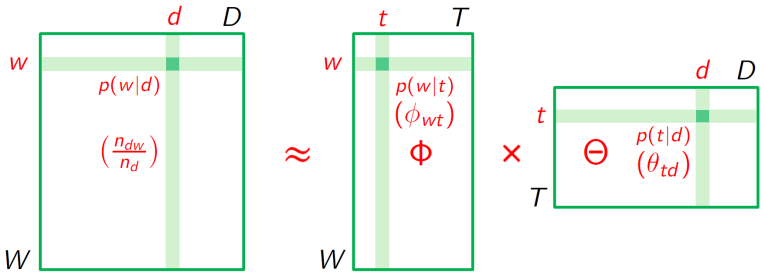


Рис. 2: Построение тематической модели сводится к вычислительной задаче поиска приближённого низкорангового стохастического матричного разложения $P = \left(\frac{n_{dw}}{n_d}\right) \approx \Phi\Theta$.

P, Φ, Θ столбцы p_d, φ_t, θ_d неотрицательны, нормированы и определяют дискретные вероятностные распределения. Такие матрицы называются *стохастическими*. Ранг произведения $\Phi\Theta$ не превышает числа тем $|T|$, которое, как правило, много меньше входных размерностей $|D|$ и $|W|$. Матрица P в общем случае имеет полный ранг, поэтому не может быть в точности равна $\Phi\Theta$. Таким образом, стохастическое матричное разложение $\Phi\Theta$ предполагается низкоранговым и приближённым.

Во-вторых, это способ мягкой би-кластеризации документов по множеству тематических кластеров T. При обычной «жёсткой» кластеризации (hard clustering) каждый документ целиком относится к одному кластеру. Тематическая модель распределяет содержимое документа d по кластерам-темам согласно распределению $p(t\,|\,d)$. Такая кластеризация называется «мягкой» (soft clustering). Аналогично, каждый терм w мягко распределяется по тем же кластерам согласно $p(t\,|\,w)$. Кластеризуются одновременно два множества, D и W. Поэтому тематическое моделирование называют *мягкой би-кластеризацией*.

В-третьих, это способ векторизации текста, т. е. получения тематических распределений для любого документа $p(t\,|\,d)$, терма $p(t\,|\,w)$, терма в контексте документа $p(t\,|\,d, w)$, текстового фрагмента или предложения $p(t\,|\,s)$. Распределение вида $p(t\,|\,x)$ называется тематическим векторным представлением, тематическим вектором, тематическим эмбедингом или *тематикой* объекта x.

В-четвёртых, это *автокодировщик* документов, состоящий из двух последовательных преобразований. *Кодировщик* $f_\Phi\colon p_d \to \theta_d$ преобразует векторное представление документа $p_d = \hat{p}(w\,|\,d)$ размерности $|W|$ в тематическое векторное представление $\theta_d = p(t\,|\,d)$ низкой размерности $|T|$. *Декодировщик* $g_\Phi\colon \theta_d \to \Phi\theta_d$ реконструирует исходное представление. Чем точнее реконструкция, тем больше информации об исходном документе удалось закодировать в векторе θ_d. В тематическом моделировании декодировщик линейный, это просто умножение вектора θ_d на матрицу Φ. Матрица Φ является параметром как кодировщика, так и декодировщика. Матрица $\Theta = (\theta_1, \ldots, \theta_D)$ является не параметром, а результатом кодирования всех документов коллекции. Это важное различие ролей двух матриц ускользает от внимания, если рассматривать тематическую модель только как матричное разложение.

В-пятых, это *языковая модель*, которая предсказывает появление термов в документах. Следует признать, что в данном качестве тематические модели довольно слабы. Чтобы хорошо предсказать терм, необходимо знать его локальный контекст, однако гипотеза «мешка слов» разрушает контексты. Это не совсем проблема, так как существуют тематические модели, использующие естественный порядок слов.

Более серьёзная проблема в том, что появление терма вряд ли определяется только тематикой терма и его контекста.

Глубокие нейронные сети на основе моделей внимания [29] и архитектуры трансформера, такие как BERT [58] или GPT [98] моделируют весь комплекс языковых явлений и предсказывают слова в тексте намного лучше, чем это делают тематические модели. И даже лучше, чем это делают люди. Модель GPT способна генерировать фейковые новости, которые люди не отличают от настоящих. В то же время, в отличие от тематических моделей, нейросетевые модели языка неинтерпретируемы. Мы пока не научились узнавать, какие именно языковые явления и каким образом смоделировала нейронная сеть. Также неясно, какой смысл имеет каждая координата нейросетевого векторного представления текста.

Итак, *целью тематического моделирования* является не столько предсказание термов в документах $p(w|d)$, сколько выявление тематической кластерной структуры текстовой коллекции; определение тематики документов $p(t|d)$ и термов $p(t|w)$; интерпретация каждой темы t средствами естественного языка.

Интерпретируемость — важнейшее свойство тематических моделей. Оно означает, что каждая тема t имеет *семантическое ядро* — подмножество термов $W_t \subset W$, которые встречаются в данной теме существенно чаще, чем во всей коллекции. Можно по-разному формализовать, что значит «существенно чаще», и даже ввести несколько взаимодополняющих пороговых критериев:

$$W_t(\alpha,\beta,\gamma) = \big\{w\colon p(w|t) \geqslant \alpha,\ p(t|w) \geqslant \beta,\ p(w|t) \geqslant \gamma p(w)\big\}.$$

Тема называется *интерпретируемой*, если эксперт или пользователь тематической модели, просмотрев список термов W_t, понимает, какой общий смысл их объединяет, может сформулировать название темы и составить её краткое изложение, пользуясь термами из ядра [45]. Данное определение не является математически строгим. Оно субъективно, так как несколько экспертов могут высказать различные мнения о теме. Однако на основе этого определения уже можно строить эмпирические методики экспертного оценивания интерпретируемости [133, 134].

Интерпретируемость тем переносится и на любые объекты x, имеющие тематический вектор $p(t|x)$. Если x является короткий фразой или нетекстовым объектом (такие задачи рассматриваются в главах 9 и 10), то для него также можно определить семантическое ядро. Согласно формуле полной вероятности,

$$W_x(\gamma) = \big\{w\colon p(w|x) = \sum_{t\in T} p(w|t)p(t|x) \geqslant \gamma p(w)\big\}.$$

Благодаря свойству интерпретируемости тема может «сама рассказать о себе» словами или термами из семантического ядра, и даже целыми фразами, отобранными из текстов методами экстрактивной суммаризации или автоматического именования тем [118]. Также и любой тематический вектор $p(t|x)$ способен рассказать о себе словами или фразами естественного языка.

Тематическая модель называется интерпретируемой, если в ней интерпретируемы все темы, за исключением, быть может, незначительной доли тем. Тема как устойчивый комплекс совместно встречающихся слов является лингвистическим явлением. Однако для его надёжного обнаружения он должен встретиться статистически значимое число раз, и для этого необходимо иметь достаточно большой объём

текстов. Интерпретируемость модели, как правило, улучшается с ростом объёма коллекции, но ухудшается с ростом числа тем. В общем случае не существует никаких математических гарантий того, что модель будет интерпретируемой. Тем не менее, интерпретируемость можно измерять и улучшать различными методами, которые рассматриваются в последующих главах этой книги.

Частотные оценки условных вероятностей. В пространстве Ω_n вероятности, выражающиеся через переменные d и w, совпадают с частотами соответствующих наблюдаемых событий:

$$p(d,w) = \frac{n_{dw}}{n}, \qquad p(d) = \frac{n_d}{n}, \qquad p(w) = \frac{n_w}{n}, \qquad p(w\,|\,d) = \frac{n_{dw}}{n_d}; \tag{3}$$

n_{dw} — число вхождений терма w в документ d;
$n_d = \sum_w n_{dw}$ — длина документа d в термах;
$n_w = \sum_d n_{dw}$ — число вхождений терма w во все документы коллекции;
$n = \sum_d \sum_w n_{dw}$ — длина коллекции в термах.

Вероятности, связанные со скрытой переменной t, тоже определяются как частоты:

$$p(t) = \frac{n_t}{n}, \qquad p(w\,|\,t) = \frac{n_{wt}}{n_t}, \qquad p(t\,|\,d) = \frac{n_{td}}{n_d}, \qquad p(t\,|\,d,w) = \frac{n_{tdw}}{n_{dw}}; \tag{4}$$

n_{tdw} — число троек, в которых терм w документа d связан с темой t;
$n_{td} = \sum_w n_{tdw}$ — число троек, в которых терм документа d связан с темой t;
$n_{wt} = \sum_d n_{tdw}$ — число троек, в которых терм w связан с темой t;
$n_t = \sum_d \sum_w n_{tdw}$ — число троек, связанных с темой t.

В отличие от (3), эти частоты не могут быть вычислены непосредственно по исходным данным, так как темы t_i неизвестны.

В вероятностном пространстве Ω выборка троек $(d_i, w_i, t_i)_{i=1}^n$ порождается из распределения $p(d,w,t)$. Поэтому формулы (3)–(4) являются не точными равенствами, а лишь *приближёнными частотными оценками* соответствующих условных вероятностей. Согласно закону больших чисел, при $n \to \infty$ частотные оценки (вероятности в пространстве Ω_n) стремятся к соответствующим вероятностям в пространстве Ω.

EM-алгоритм. Заметим, что все оценки (4) выражаются через $n_{tdw} = n_{dw} p(t\,|\,d,w)$. Зная условные распределения $p(t\,|\,d,w)$, можно оценить искомые параметры тематической модели $\varphi_{wt} = p(w\,|\,t)$ и $\theta_{td} = p(t\,|\,d)$. И, наоборот, зная параметры модели, можно выразить условные вероятности $p(t\,|\,d,w)$ по формуле Байеса:

$$p(t\,|\,d,w) = \frac{p(t,w\,|\,d)}{p(w\,|\,d)} = \frac{p(w\,|\,t)p(t\,|\,d)}{p(w\,|\,d)} = \frac{\varphi_{wt}\theta_{td}}{\sum_s \varphi_{ws}\theta_{sd}}. \tag{5}$$

Алгоритм 2. Рациональный EM-алгоритм для тематической модели (2).

Вход: коллекция D, число тем $|T|$, число итераций $i_{\max}$;
Выход: матрицы термов тем Φ и тем документов Θ;

```
инициализация φ_wt, θ_td для всех d ∈ D, w ∈ W, t ∈ T;
для всех итераций i = 1, ..., i_max
    обнулить n_wt, n_td, n_t для всех d ∈ D, w ∈ W, t ∈ T;
    для всех документов d ∈ D и всех термов w ∈ d
        n_tdw := n_dw φ_wt θ_td / Σ_z φ_wz θ_zd для всех t ∈ T;
        увеличить n_wt, n_td, n_t на n_tdw для всех t ∈ T;
    φ_wt := n_wt / n_t для всех w ∈ W, t ∈ T;
    θ_td := n_td / n_d для всех d ∈ D, t ∈ T;
```

Таким образом, получаем систему нелинейных уравнений относительно параметров модели φ_{wt}, θ_{td} и вспомогательных переменных p_{tdw}, n_{wt}, n_{td}, n_t:

$$p_{tdw} = \frac{\varphi_{wt}\theta_{td}}{\sum_s \varphi_{ws}\theta_{sd}}; \tag{6}$$

$$\varphi_{wt} = \frac{n_{wt}}{n_t}; \qquad n_{wt} = \sum_{d\in D} n_{dw}p_{tdw}; \qquad n_t = \sum_{w\in W} n_{wt}; \tag{7}$$

$$\theta_{td} = \frac{n_{td}}{n_d}; \qquad n_{td} = \sum_{w\in d} n_{dw}p_{tdw}. \tag{8}$$

Для решения данной системы уравнений удобно применить метод простых итераций: сначала выбрать начальные приближения параметров φ_{wt} и θ_{td}, по ним вычислить вспомогательные переменные p_{tdw} с помощью которых найти следующее приближение параметров φ_{wt} и θ_{td}. Вычисления по формулам (6)–(8) продолжаются в цикле до сходимости.

Этот итерационный процесс является частным случаем EM-алгоритма [57]. Вычисление условных распределений скрытых переменных (6) называется E-шагом (expectation), вычисление параметров модели (7)–(8) — M-шагом (maximization).

Далее мы выведем этот алгоритм из общей оптимизационной постановки задачи. Сейчас мы пришли к нему элементарным путём, который даёт интуитивно понятные формулы, но оставляет без ответов важные вопросы: сходится ли алгоритм к решению системы уравнений, единственно ли это решение, и почему эта система описывает тематическую модель, приближающую $\hat{p}(w\,|\,d)$.

Рациональный EM-алгоритм. Вычисление переменных n_{wt}, n_{td}, n_t на M-шаге требует однократного прохода коллекции в цикле по всем термам $w \in d$ всех документов $d \in D$. Внутри этого цикла значение p_{tdw} вычисляется только один раз. Поэтому E-шаг можно встроить внутрь M-шага без дополнительных вычислительных затрат и без хранения трёхмерной матрицы p_{tdw}. Этот вариант EM-алгоритма будем называть *рациональным*; его псевдокод показан в алгоритме 2.

Выводы по главе

- Задача тематического моделирования состоит в том, чтобы по заданной коллекции текстов найти семантические ядра тем $p(w|t)$ и тематические векторные представления документов $p(t|d)$.
- Тематическая модель — это и низкоранговое стохастическое матричное разложение, и мягкая би-кластеризация, и вероятностная векторизация текстов, и автокодировщик, и языковая модель.
- Основная цель тематического моделирования — выяснить, «о чём все эти тексты», не читая их. Для этого темы должны обладать свойством интерпретируемости. Многие виды тематических моделей направлены на улучшение интерпретируемости тем.
- Никому не нравится гипотеза «мешка слов» — ведь текст теряет смысл, если в нём перепутать слова. Можно ли от неё отказаться? Да, и многими способами, дойдём до этого в главе 15.
- Все введённые обозначения будут активно использоваться далее.
- EM-алгоритм будет обоснован в двух следующих главах.

3 Максимизация на единичных симплексах

В этой главе будет доказана лемма о максимизации гладкой функции на единичных симплексах. В дальнейшем с её помощью будут выводиться формулы EM-подобных итерационных процессов для любых тематических моделей, от самых простых до самых сложных. Применимость этой леммы выходит далеко за пределы тематического моделирования. Она позволяет оптимизировать любые модели, параметрами которых являются неотрицательные нормированные векторы.

Необходимые условия максимума. Классическим инструментом решения задач оптимизации с ограничениями равенствами и неравенствами являются *условия Каруша–Куна–Таккера*. Рассмотрим задачу математического программирования

$$\begin{cases} f(x) \to \max\limits_x; \\ g_i(x) \leqslant 0, \quad i = 1, \ldots, m; \\ h_j(x) = 0, \quad j = 1, \ldots, k. \end{cases}$$

Теорема 3.1. *Пусть функции $f(x)$, $g_i(x)$, $h_j(x)$ непрерывно дифференцируемы в точке x. Если x — точка локального максимума и задача удовлетворяет условиям регулярности, то существуют значения μ_i, $i = 1, \ldots, m$, λ_j, $j = 1, \ldots, k$, называемые множителями Лагранжа, такие, что:*

$$\begin{cases} \dfrac{\partial \mathscr{L}}{\partial x} = 0, \quad \mathscr{L}(x; \mu, \lambda) = f(x) - \sum\limits_{i=1}^{m} \mu_i g_i(x) - \sum\limits_{j=1}^{k} \lambda_j h_j(x); \\ g_i(x) \leqslant 0;\ h_j(x) = 0;\ \textit{(исходные ограничения)} \\ \mu_i \geqslant 0;\ \textit{(двойственные ограничения)} \\ \mu_i g_i(x) = 0;\ \textit{(условие дополняющей нежёсткости)} \end{cases} \tag{9}$$

Условия регулярности могут формулироваться по-разному. В задачах с линейными ограничениями они выполнены всегда.

Задача (15), (14) относится к классу невыпуклых задач математического программирования. Для неё возможно найти лишь локальный экстремум, качество которого будет зависеть от начального приближения. Далее мы не будем уделять много внимания вопросам инициализации, сосредоточившись на модификациях самой функции f, которые способны влиять на решение гораздо сильнее.

Максимизация гладкой функции на единичных симплексах. Введём оператор norm, который преобразует произвольный вектор $(x_i)_{i \in I}$ в вектор вероятностей $(p_i)_{i \in I}$ дискретного распределения путём обнуления отрицательных элементов и нормировки:

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

где $(x)_+ = \max\{0, x\}$ — операция положительной срезки. Если $x_i \leqslant 0$ для всех $i \in I$, то результатом оператора norm по определению является нулевой вектор.

Лемма 3.2 (о максимизации на единичных симплексах). *Пусть функция $f(\Omega)$ непрерывно дифференцируема по набору векторов $\Omega = (\omega_j)_{j\in J}$, $\omega_j = (\omega_{ij})_{i\in I_j}$, вообще говоря, различных размерностей $|I_j|$. Тогда векторы ω_j локального экстремума задачи*

$$\begin{cases} f(\Omega) \to \max\limits_{\Omega}; \\ \sum\limits_{i\in I_j} \omega_{ij} = 1, \quad j \in J; \\ \omega_{ij} \geqslant 0, \quad i \in I_j,\ j \in J; \end{cases}$$

при условии 1^0 : $(\exists i \in I_j)\ \omega_{ij}\frac{\partial f}{\partial \omega_{ij}} > 0$ удовлетворяют уравнениям

$$\omega_{ij} = \operatorname*{norm}_{i\in I_j}\left(\omega_{ij}\frac{\partial f}{\partial \omega_{ij}}\right), \quad i \in I_j; \tag{10}$$

при условии 2^0 : $(\forall i \in I_j)\ \omega_{ij}\frac{\partial f}{\partial \omega_{ij}} \leqslant 0$ и $(\exists i \in I_j)\ \omega_{ij}\frac{\partial f}{\partial \omega_{ij}} < 0$ они удовлетворяют уравнениям

$$\omega_{ij} = \operatorname*{norm}_{i\in I_j}\left(-\omega_{ij}\frac{\partial f}{\partial \omega_{ij}}\right), \quad i \in I_j;$$

в противном случае (условие 3^0) они удовлетворяют однородным уравнениям $\omega_{ij}\frac{\partial f}{\partial \omega_{ij}} = 0$, $i \in I_j$.

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} — ограничениям неотрицательности. Запишем условия Каруша–Куна–Таккера (9), приравняв нулю производные лагранжиана по параметрам модели:

$$\frac{\partial \mathscr{L}}{\partial \omega_{ij}} = \frac{\partial f}{\partial \omega_{ij}} - \lambda_i + \mu_{ij} = 0; \quad \mu_{ij}\omega_{ij} = 0. \tag{11}$$

Зафиксируем $j \in J$. Предположим, что $\omega_{ij} > 0$. Тогда $\mu_{ij} = 0$. Умножив обе части равенства (11) на ω_{ij}, получим уравнение

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

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

Возможны три случая.

1. Пусть существует индекс $k \in I_j$ такой, что $A_{kj} > 0$. Тогда $\lambda_j > 0$. Если $A_{ij} \leqslant 0$ при некотором $i \in I_j$, то уравнение не может быть выполнено, и полученное противоречие означает, что сделанное вначале предположение $\omega_{ij} > 0$ не верно, 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 левую и правую части этого уравнения по i, выразим двойственную переменную: $\lambda_j = \sum_{i \in I_j}(A_{ij})_+$. Подставляя полученное значение λ_j в формулу $\omega_{ij} = \frac{1}{\lambda_j}(A_{ij})_+$, получим искомое уравнение (10).

2. Если $A_{ij} \leqslant 0$ для всех $i \in I_j$ и хотя бы одно из этих неравенств строгое, то $\lambda_j < 0$, и аналогичным образом мы получаем второе искомое уравнение.

3. Если $A_{ij} = 0$ для всех $i \in I_j$, то $\lambda_j = 0$, и вектор ω_{ij} определяется из системы однородных уравнений $\omega_{ij}\frac{\partial f}{\partial \omega_{ij}} = 0$ при условиях неотрицательности и нормировки.

Лемма доказана.

Для решения системы уравнений удобно использовать метод простой итерации. Сначала задаются начальные приближения для векторов ω_j, затем их значения многократно обновляются по формуле (10). Итерационный процесс похож на обычный градиентный метод безусловной максимизации

$$\omega_{ij} = \omega_{ij} + \eta\frac{\partial f}{\partial \omega_{ij}},$$

отличаясь от него отсутствием параметра градиентного шага η и наличием операции norm для проецирования полученного вектора на единичный симплекс.

Лемма о максимизации на единичных симплексах может применяться не только в тематическом моделировании, но и для оптимизации любых моделей, параметрами которых являются дискретные вероятностные распределения или неотрицательные нормированные векторы. В частности, её можно использовать для обучения нейронных сетей, в которых такие ограничения накладываются на некоторые векторы параметров. При этом градиенты $\frac{\partial f}{\partial \omega_i}$ можно по-прежнему вычислять методом обратного распространения ошибок. Ограничения приводят к модификации градиентного шага, но никак не сказываются на вычислении самих градиентов.

Сходимость итерационного процесса. Рассмотрим итерационный процесс, основанный на формуле (10), при некотором начальном приближении $\Omega^0 = (\omega_j^0)_{j \in J}$:

$$\omega_{ij}^{t+1} = \underset{i \in I_j}{\operatorname{norm}}\left(\omega_{ij}^t \frac{\partial f(\Omega^t)}{\partial \omega_{ij}^t}\right), \quad t = 0, 1, 2, \ldots \tag{12}$$

Утверждение 3.3. *Пусть $f(\Omega)$ — ограниченная сверху, непрерывно дифференцируемая функция, и все Ω^t, начиная с некоторой итерации t^0 обладают свойствами:*

1) $\forall j \in J \;\; \forall i \in I_j \;\; \omega_{ij}^t = 0 \to \omega_{ij}^{t+1} = 0$ *(сохранение нулей);*

2) $\exists \epsilon > 0 \;\; \forall j \in J \;\; \forall i \in I_j \;\; \omega_{ij}^t \notin (0, \epsilon)$ *(отделимость от нуля);*

3) $\exists \delta > 0 \;\; \forall j \in J \;\; \exists i \in I_j \;\; \omega_{ij}^t \frac{\partial f(\Omega^t)}{\partial \omega_{ij}} \geqslant \delta$ *(невырожденность);*

Тогда $\left|\omega_{ij}^{t+1} - \omega_{ij}^t\right| \to 0$ *при* $t \to \infty$.

Данное утверждение является обобщением результатов о сходимости EM-алгоритма в тематическом моделировании [12] и приводится здесь без доказательства. В условиях Теоремы 3.3 все предельные точки последовательности $\{\Omega^t\}_{t=0}^{\infty}$ являются неподвижными точками преобразования (12). На практике свойства 1–3 могут быть гарантированы при реализации итерационного алгоритма.

Выводы по главе

- Лемма о максимизации гладкой функции на единичных симплексах будет использоваться далее для вывода всех без исключения алгоритмов тематического моделирования.
- Такой способ намного проще байесовского вывода, который обычно применяется в литературе по тематическому моделированию. Сравнение этих подходов отложим до главы 7.
- На практике будет использоваться только первое условие (10).
- Сходство этого условия с градиентным шагом позволяет обучать нейронные сети с неотрицательными нормированными весами. Интересная возможность, но за рамками данной книги.
- Какова связь доказанной леммы с EM-алгоритмом из предыдущей главы, узнаем в следующей главе.

4 Аддитивная регуляризация

В этой главе мы введём общий формализм аддитивной регуляризации, который позволяет наделять тематические модели разнообразными полезными свойствами. Что это за свойства, и как они формализуются с помощью регуляризации, будем разбираться в следующих главах. Здесь мы поставим задачу тематического моделирования как задачу максимизации на единичных симплексах. С помощью леммы 3.2 выведем EM-алгоритм буквально в две строчки.

Принцип максимума правдоподобия используется в математической статистике для оценивания неизвестных параметров вероятностных моделей по наблюдаемым данным. Согласно этому принципу, выбираются такие значения параметров, при которых наблюдаемая выборка наиболее правдоподобна.

Функция правдоподобия определяется как зависимость вероятности наблюдаемой выборки $X = (d_i, w_i)_{i=1}^n$ от параметров модели Φ, Θ. В силу гипотезы о независимости элементов выборки она равна произведению вероятностей термов в документах:

$$p\big(X;\Phi,\Theta\big) = \prod_{i=1}^{n} p(d_i, w_i) = \prod_{d\in D}\prod_{w\in d} p(w\,|\,d)^{n_{dw}} \underbrace{p(d)^{n_{dw}}}_{\text{const}} \;\to \max_{\Phi,\Theta}.$$

Прологарифмировав правдоподобие, перейдём от произведения к сумме и отбросим слагаемые, не зависящие от параметров модели. Получим задачу максимизации логарифма правдоподобия при ограничениях неотрицательности и нормировки:

$$\sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{t\in T} \varphi_{wt}\theta_{td} \;\to\; \max_{\Phi,\Theta}; \tag{13}$$

$$\sum_{w\in W} \varphi_{wt} = 1; \quad \varphi_{wt} \geqslant 0; \qquad \sum_{t\in T} \theta_{td} = 1; \quad \theta_{td} \geqslant 0. \tag{14}$$

Регуляризация некорректно поставленных задач. Задача называется *корректно поставленной* по Адамару, если её решение существует, единственно и устойчиво.

Задача стохастического матричного разложения является некорректно поставленной, так как множество её решений в общем случае бесконечно. Если $\Phi\Theta$ — решение, то $(\Phi S)(S^{-1}\Theta)$ также является решением для всех невырожденных матриц S, при условии, что матрицы ΦS и $S^{-1}\Theta$ — стохастические.

Существует общий подход к решению некорректно поставленных обратных задач, называемый *регуляризацией* [17]. Когда оптимизационная задача недоопределена, к основному критерию добавляют дополнительный критерий — регуляризатор, учитывающий специфику решаемой задачи и знания предметной области. В практических задачах автоматической обработки текстов дополнительных критериев и ограничений на решение может быть много.

Аддитивная регуляризация тематических моделей (ARTM) [5] основана на максимизации линейной комбинации логарифма правдоподобия и *регуляризаторов* $R_i(\Phi,\Theta)$ с неотрицательными *коэффициентами регуляризации* τ_i, $i = 1,\ldots,k$:

$$\sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{t\in T} \varphi_{wt}\theta_{td} + R(\Phi,\Theta) \;\to\; \max_{\Phi,\Theta}; \qquad R(\Phi,\Theta) = \sum_{i=1}^{k} \tau_i R_i(\Phi,\Theta); \tag{15}$$

при ограничениях неотрицательности и нормировки (14).

Преобразование вектора критериев в один скалярный критерий — это один из базовых приёмов в многокритериальной оптимизации, называемый *скаляризацией*.

Основная теорема ARTM. Применим лемму 3.2 к набору вектор-столбцов двух матриц $\Omega = (\Phi, \Theta)$. В дальнейшем эта лемма пригодится нам и в более сложных случаях, когда матриц будет больше двух.

Теорема 4.1. *Пусть функция $R(\Phi, \Theta)$ непрерывно дифференцируема. Тогда точка (Φ, Θ) локального экстремума задачи (15) с ограничениями (14) удовлетворяет системе уравнений со вспомогательными переменными $p_{tdw} = p(t\,|\,d, w)$, если из решения исключить нулевые столбцы матриц Φ, Θ:*

$$p_{tdw} = \underset{t\in T}{\operatorname{norm}}\big(\varphi_{wt}\theta_{td}\big); \tag{16}$$

$$\varphi_{wt} = \underset{w\in W}{\operatorname{norm}}\left(n_{wt} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}}\right); \qquad n_{wt} = \sum_{d\in D} n_{dw}p_{tdw}; \tag{17}$$

$$\theta_{td} = \underset{t\in T}{\operatorname{norm}}\left(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right); \qquad n_{td} = \sum_{w\in d} n_{dw}p_{tdw}. \tag{18}$$

Доказательство. Перепишем (16) 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 (15) основное решение (10) из леммы 3.2 и выделим в полученных выражениях вспомогательные переменные p_{tdw}:

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

Нулевые столбцы в матрицах Φ и Θ получаются в тех случаях, когда в лемме 3.2 вместо условия 1^0 реализуются условия 2^0 или 3^0. В таких случаях основное решение (10) уже неприменимо, однако теорема исключает их из рассмотрения.

Теорема доказана.

Следствие 4.2. *В условиях теоремы 4.1 точка локального экстремума (Φ, Θ) оптимизационной задачи (15) с ограничениями (14) совпадает с точкой локального экстремума следующей оптимизационной задачи с ограничениями (14) и (16):*

$$\sum_{d\in D}\sum_{w\in W}\sum_{t\in T} n_{dw}p_{tdw}\ln\big(\varphi_{wt}\theta_{td}\big) + R(\Phi,\Theta) \;\to\; \max_{\Phi,\Theta}. \tag{19}$$

Доказательство. Представив в (19) логарифм произведения суммой двух логарифмов, получим две суммы, первая зависит только от Φ, вторая — только от Θ:

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

Применяя к этой задаче формулу (10) из леммы 3.2 при ограничениях (14) и фиксированных значениях переменных p_{tdw}, немедленно получим, что необходимое условие локального экстремума в задаче (19) эквивалентно уравнениям (17)–(18).

Следствие доказано.

Таким образом, M-шаг — это решение задачи максимизации правдоподобия для определения основных параметров модели Φ, Θ при известных вспомогательных переменных p_{tdw}. В специальных случаях такое промежуточное представление EM-алгоритма оказывается более удобным для вывода формул M-шага.

Регуляризованный EM-алгоритм. Систему (16)–(18) удобно решать методом простых итераций. Сначала выбираются начальные приближения $\varphi_{wt}, \theta_{td}$, затем в цикле до сходимости чередуются *E-шаг* (16) и *M-шаг* (17)–(18). Каждая итерация представляет собой проход по всем термам всех документов коллекции.

Рациональный EM-алгоритм с регуляризацией строится аналогично алгоритму 2, только шаги 7 и 8 заменяются формулами M-шага (17)–(18).

В алгоритме 3 показано ещё несколько усовершенствований:

1) вектор θ_d определяется по окончании обработки документа d, а не всей коллекции, поэтому далее он может не храниться, как и счётчики n_{td};

2) при обработке документа d может быть сделано несколько итераций по всем термам документа $w \in d$ до сходимости вектора θ_d;

3) вектор θ_d инициализируется равномерным распределением;

4) вектор φ_t инициализируется случайным распределением, которое отделяется от нуля распределением $p_w = n_w/n$;

Известно, что EM-алгоритм без регуляризации сходится в слабом смысле: на каждой итерации правдоподобие увеличивается [57]. Аналогичные условия слабой сходимости для ARTM получены в [12].

Онлайновый EM-алгоритм считается наиболее быстрым и хорошо распараллеливается [76, 33]. Обновления матрицы Φ производятся не после прохода всей коллекции, а чаще — после обработки определённого объёма данных. Если коллекция большая, то матрица Φ может сойтись и перестать меняться задолго до окончания первой итерации. Тогда одного прохода будет достаточно для построения модели. Поэтому онлайновые алгоритмы способны обрабатывать потоковые данные.

Детали параллельной реализации оффлайнового и онлайнового EM-алгоритма в библиотеке `BigARTM` описаны в главах 19 и 20, ещё подробнее — в статьях [68, 3].

Условия вырожденности. В тематическом моделировании случаи 2^0 и 3^0 из леммы 3.2 будем считать вырожденными. Если нормировочный знаменатель в (10) окажется равным нулю, то соответствующий вектор ω_j будем полагать нулевым и исключать его из модели, сокращая размерность пространства параметров и считая это полезным побочным действием регуляризации. Если вырожденность нежелательна, то её нетрудно избежать, уменьшая коэффициент регуляризации.

Алгоритм 3. Регуляризованный ЕМ-алгоритм для ARTM.

Вход: коллекция D, число тем $|T|$, число итераций $i_{\max}$;
Выход: матрицы термов тем Φ и тем документов Θ;

инициализация $\varphi_{wt} := \underset{w\in W}{\operatorname{norm}}(p_w + \operatorname{rand})$ для всех $w \in W$, $t \in T$;
для всех итераций $i = 1, \ldots, i_{\max}$
  $n_{wt} := 0$ для всех $w \in W$, $t \in T$;
  для всех документов $d \in D$
    инициализация $\theta_{td} := \frac{1}{|T|}$ для всех $t \in T$;
    повторять
      $n_{tdw} := n_{dw} \underset{t\in T}{\operatorname{norm}}\big(\varphi_{wt}\theta_{td}\big)$ для всех $w \in d,\ t \in T$;
      $\theta_{td} := \underset{t\in T}{\operatorname{norm}}\Big(\sum\limits_{w\in d} n_{tdw} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\Big)$ для всех $t \in T$;
    пока θ_d не сойдётся;
    $n_{wt} := n_{wt} + n_{tdw}$ для всех $w \in d,\ t \in T$;
  $\varphi_{wt} := \underset{w\in W}{\operatorname{norm}}\Big(n_{wt} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}}\Big)$ для всех $w \in W$, $t \in T$;

Тема t называется *вырожденной*, если

$$n_{wt} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}} \leqslant 0 \text{ для всех } w \in W.$$

Вырожденность является следствием сильного разреживающего воздействия регуляризатора R. Обнуление столбца матрицы Φ означает, что для максимизации регуляризатора выгодно исключить данную тему из модели.

Документ d называется *вырожденным*, если

$$n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}} \leqslant 0 \text{ для всех } t \in T.$$

Вырожденность документа может означать, что модель не в состоянии его описать, например, если он слишком короткий или не соответствует тематике коллекции. Обнуление столбца матрицы Θ означает, что, для максимизации регуляризатора данный документ выгодно считать аномальным.

Вероятностный латентный семантический анализ (probabilistic latent semantic analysis, PLSA) — это исторически первая вероятностная тематическая модель. Она была предложена Томасом Хофманном в 1999 году [77]. В ARTM ей соответствует нулевой регуляризатор $R(\Phi, \Theta) = 0$. В этом случае система (16)–(18) совпадает с системой (6)–(8), которую мы получили из простых эвристических соображений. Добавление регуляризации меняет только формулы М-шага, не затрагивая Е-шаг.

Частотные оценки условных вероятностей $\varphi^*_{wt} = \frac{n_{wt}}{n_t}$ и $\theta^*_{td} = \frac{n_{td}}{n_d}$ в модели PLSA называются *несмещёнными оценками* максимального правдоподобия, в отличие от *смещённых оценок* в моделях с ненулевым регуляризатором.

В модели PLSA не может быть вырожденных тем или документов, поскольку вырожденность возникает при отрицательных производных регуляризатора.

Улучшение сходимости. Метод простой итерации в общем случае не гарантирует сходимости к стационарной точке регуляризованного правдоподобия [12]. Этот недостаток исправляется простой модификацией, не требующей дополнительных затрат времени или памяти. Достаточно подставить в формулы M-шага несмещённые частотные оценки $\varphi^*_{wt} = \frac{n_{wt}}{n_t}$ и $\theta^*_{td} = \frac{n_{td}}{n_d}$ вместо текущих значений φ_{wt} и θ_{td}:

$$\varphi_{wt} = \underset{w \in W}{\operatorname{norm}} \left(n_{wt} + \varphi^*_{wt} \frac{\partial R(\Phi^*, \Theta^*)}{\partial \varphi^*_{wt}} \right); \tag{20}$$

$$\theta_{td} = \underset{t \in T}{\operatorname{norm}} \left(n_{td} + \theta^*_{td} \frac{\partial R(\Phi^*, \Theta^*)}{\partial \theta^*_{td}} \right). \tag{21}$$

Эксперименты в [12] показали, что данная модификация приводит к более высоким значениям регуляризованного правдоподобия, причём итерационный процесс гораздо быстрее (уже на второй итерации) входит в режим монотонного увеличения регуляризованного правдоподобия. Чем больше коэффициент регуляризации, тем сильнее проявляется эффект улучшения сходимости.

О стратегиях регуляризации. Задача тематического моделирования по сути является многокритериальной. Темы должны удовлетворять многим требованиям одновременно: интерпретируемости, различности, разреженности и т. д. Тематическая модель обычно используется как вспомогательный инструмент для решения одной или сразу нескольких задач текстовой аналитики — информационного поиска, визуализации, категоризации, сегментации, суммаризации и т. д. Каждая задача предъявляет свои требования к модели. В ARTM все требования формализуются в виде критериев регуляризации R_i и балансируются с помощью коэффициентов τ_i. Коэффициенты τ_i приходится подбирать в каждой задаче экспериментально, чтобы найти компромисс между всеми критериями. Более того, для измерения качества модели обычно используются не сами регуляризаторы R_i, а какие-то другие *метрики качества*. Регуляризаторы должны быть гладкими функциями, удобными для вычислений на M-шаге. Метрики качества должны иметь удобные для интерпретации числовые значения. К сожалению, эти требования часто входят в противоречие. Например, общепринятые метрики качества информационного поиска почти никогда не являются гладкими функциями.

На практике проблема выбора коэффициентов регуляризации перерастает в более общую проблему *управления качеством* модели путём изменения коэффициентов τ_i в ходе итераций. Регуляризаторы можно вводить в итерационный процесс постепенно, по одному. При этом каждый регуляризатор будет подготавливать модель к воздействию последующих. Некоторые регуляризаторы рекомендуется включать лишь после того, как EM-алгоритм начал сходиться. Другие лучше отключать после того, как они оказали необходимое воздействие на модель. Некоторые регуляризаторы могут нейтрализовать друг друга, тогда их приходится чередовать.

Образно говоря, регуляризаторы подобны лекарствам для модели — в малых дозах лечат, в больших смертельно опасны, в сочетаниях могут давать неожиданные эффекты. Систематизация этих эффектов становится предметом эмпирических исследований в ARTM.

Стратегией регуляризации будем называть правила изменения коэффициентов регуляризации τ_i в ходе итераций EM-алгоритма. Эти правила могут использовать текущие значения метрик качества и параметров модели.

Общий подход к автоматической настройке коэффициентов регуляризации и других гиперпараметров для ARTM был предложен в [90] на основе эволюционного алгоритма и представления процесса обучения как многоэтапной стратегии изменения гиперпараметров. В последующей работе [89] этот метод был дополнен суррогатной моделью оценивания качества, что позволило примерно в 1,5 раза сократить время поиска гиперпараметров. В этих работах использовался обычный для *автоматического машинного обучения* (AutoML) подход: для каждой точки в пространстве гиперпараметров модель обучается заново по обучающей выборке, затем её качество оценивается по тестовой выборке. На больших данных такая стратегия представляется слишком затратной. Корректировать гиперпараметры можно и чаще — обработав лишь часть коллекции и не дожидаясь окончательной сходимости параметров модели. Кроме того, для оценивания качества моделей можно использовать несколько различных метрик, дополняющих друг друга. Построение быстрых многокритериальных стратегий регуляризации пока остаётся открытой проблемой.

Относительные коэффициенты регуляризации. Коэффициенты регуляризации, тщательно подобранные для одной коллекции, могут плохо подходить для другой. Они могут зависеть от размера коллекции, мощности словаря, средней длины документов. Если коллекция пополняется, то со временем может потребоваться их перенастройка. Проблема решается с помощью нормировки и введения *относительных коэффициентов регуляризации*, выражающих степень воздействия регуляризатора на тематическую модель. Относительные коэффициенты могут оставаться фиксированными по мере роста коллекции, и для них могут оцениваться универсальные рекомендуемые значения, подходящие для любых задач.

Рассмотрим формулу M-шага (17) со взвешенной суммой регуляризаторов R_i:

$$\varphi_{wt} = \underset{w\in W}{\operatorname{norm}}\left(n_{wt} + \sum_{i=1}^{k} \tau_i \varphi_{wt} \frac{\partial R_i}{\partial \varphi_{wt}}\right).$$

Введём *суммарное воздействие* r_{it} регуляризатора R_i на тему t и его *суммарное воздействие* r_i на все темы:

$$r_{it} = \sum_{w\in W} \left|\varphi_{wt} \frac{\partial R_i}{\partial \varphi_{wt}}\right|, \qquad r_i = \sum_{t\in T} r_{it}.$$

Введение нормировки $\tau_i = \tilde{\tau}_i \frac{n}{r_i}$ позволяет интерпретировать коэффициент $\tilde{\tau}_i$ как *относительное воздействие* регуляризатора R_i на тематическую модель. Он показывает, во сколько раз влияние регуляризатора на модель сильнее влияния исходных данных. При $\tilde{\tau}_i \to 0$ регуляризатор R_i отключается. При $\tilde{\tau}_i \to \infty$ перерегуляризация может приводить к вырождению модели.

Введение другой нормировки $\tau_i = \tilde{\tau}_i \frac{n_t}{r_{it}}$ позволяет интерпретировать коэффициент $\tilde{\tau}_i$ как *относительное воздействие* регуляризатора R_i на отдельную тему t. Теперь абсолютный коэффициент регуляризации τ_i становится зависящим от темы, однако его относительные воздействия на все темы одинаковы.

В общем случае не известно, какая из двух нормировок лучше. Для общности введём выпуклую комбинацию двух нормировок:

$$\tau_i = \tilde{\tau}_i \left(\gamma_i \frac{n_t}{r_{it}} + (1-\gamma_i)\frac{n}{r_i}\right),$$

где $\tilde{\tau}_i$ — *относительный коэффициент регуляризации*; параметр γ_i назовём *степенью индивидуализации* воздействия регуляризатора R_i на темы. При $\gamma_i = 1$ коэффициенты τ_i максимально различаются по темам, выравнивая относительные воздействия регуляризатора R_i на темы. При $\gamma_i = 0$ коэффициенты τ_i не различаются по темам. Параметр γ_i предлагается подбирать экспериментальным путём.

Аналогично рассмотрим формулу М-шага (18) со взвешенной суммой регуляризаторов R_i. Введём *суммарное воздействие* q_{id} регуляризатора R_i на документ d и его *суммарное воздействие* q_i на коллекцию:

$$q_{id} = \sum_{t \in T} \left| \theta_{td} \frac{\partial R_i}{\partial \theta_{td}} \right|, \qquad q_i = \sum_{d \in D} q_{id}.$$

Представим коэффициент регуляризации τ_i в виде

$$\tau_i = \tilde{\tau}_i \left(\gamma_i \frac{n_d}{q_{id}} + (1 - \gamma_i) \frac{n}{q_i} \right),$$

где $\tilde{\tau}_i$ — *относительный коэффициент регуляризации*, γ_i — *степень индивидуализации* воздействия регуляризатора R_i на документы. При $\gamma_i = 1$ коэффициенты максимально различаются по документам, выравнивая относительные воздействия регуляризатора на документы. Высокая индивидуализация полезна, когда документы коллекции существенно различаются по длинам n_d. При $\gamma_i = 0$ коэффициенты не различаются по документам.

Выводы по главе

- Наконец мы обосновали EM-алгоритм, полученный в первой главе из элементарных соображений.
- Более того, мы его обогатили введением регуляризаторов, оставив полную свободу дальнейшего выбора функций $R_i(\Phi, \Theta)$.
- Заодно обсудили ещё несколько аспектов применения ARTM: инициализацию, вырожденность, сходимость, стратегии регуляризации, относительные коэффициенты регуляризации.
- Данная глава играет ключевую роль в книге. Дальнейшее изложение будет развитием теории аддитивной регуляризации.
- Применение основной леммы о максимизации на единичных симплексах будет оставаться столь же простым, даже когда модели будут иметь более сложную структуру.

5 Вероятностная регуляризация и модель LDA

Модель латентного размещения Дирихле (latent Dirichlet allocation) LDA [40, 41] является, пожалуй, наиболее цитируемой в тематическом моделировании. В популярных обзорах её иногда отождествляют со всем тематическим моделированием, хотя в литературе можно найти сотни других моделей. Своей публикацией 2003 года Дэвид Блэй, Эндрю Ын и Майкл Джордан определили долгосрочный тренд развития тематического моделирования в рамках байесовского обучения.

В данной главе мы дадим более простое обоснование LDA с позиций аддитивной регуляризации. Для этого нам не понадобятся ни априорные распределения Дирихле, ни сложный математический аппарат байесовского вывода, рассмотрение которого мы отложим до главы 7.

Принцип максимума апостериорной вероятности. До сих пор мы предполагали, что данные порождаются вероятностной моделью с параметрами (Φ, Θ), которые не известны и не случайны. Теперь предположим, что параметры сами являются случайными переменными и подчиняются *априорному распределению* $p(\Phi, \Theta; \gamma)$ с неслучайным *вектором гиперпараметров* γ. В этом случае максимизация совместного правдоподобия данных $X = (d_i, w_i)_{i=1}^n$ и модели (Φ, Θ) приводит к принципу *максимума апостериорной вероятности* (maximum a posteriori probability, MAP):

$$p(X, \Phi, \Theta; \gamma) = p(X \,|\, \Phi, \Theta)\, p(\Phi, \Theta; \gamma) = p(\Phi, \Theta; \gamma) \prod_{i=1}^{n} p(d_i, w_i \,|\, \Phi, \Theta) \;\to\; \max_{\Phi, \Theta, \gamma}.$$

После логарифмирования получаем модификацию задачи (13), в которой логарифм априорного распределения становится регуляризатором:

$$\sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \varphi_{wt} \theta_{td} + \underbrace{\ln p(\Phi, \Theta; \gamma)}_{R(\Phi, \Theta)} \;\to\; \max_{\Phi, \Theta, \gamma}. \tag{22}$$

Во многих случаях возможен и обратный переход. Регуляризатор $R(\Phi, \Theta)$, изначально не имеющий вероятностной интерпретации, можно преобразовать в априорное распределение путём эспоненцирования и нормировки: $p(\Phi, \Theta) \propto \exp(R(\Phi, \Theta))$. Нормировочный множитель не зависит от параметров модели, поэтому для максимизации апостериорной вероятности по (Φ, Θ) он не важен; его отсутствие можно компенсировать подбором коэффициента регуляризации. Однако если решать задачу оптимизации гиперпараметров γ, то его уже придётся вычислять и учитывать.

Априорные распределения Дирихле. Основной мотивацией для введения модели LDA в [41] было решение проблемы переобучения в модели PLSA. Проблема заключалась в том, что модель PLSA предсказывала вероятности термов $p(w \,|\, d)$ на новых документах заметно хуже, чем на обучающей коллекции. Обычно переобучение связано с избыточной размерностью пространства параметров, поэтому на матрицы Φ, Θ следует накладывать дополнительные ограничения. В [41] было введено предположение, что столбцы этих матриц являются случайными векторами и порождаются

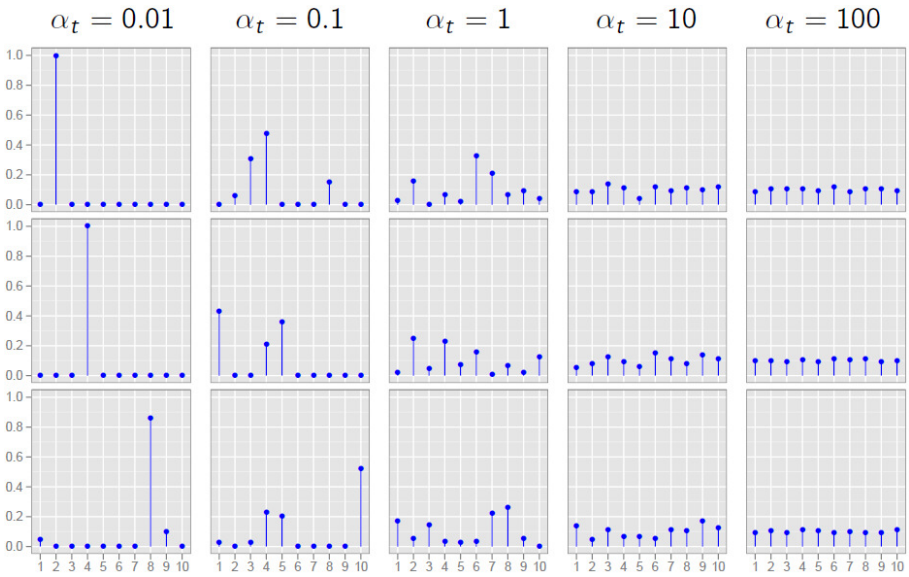


Рис. 3: Неотрицательные нормированные векторы θ размерности $|T| = 10$, порождённые симметричными распределениями Дирихле, по три вектора для каждого из пяти значений параметра α_t.

распределениями Дирихле с гиперпараметрами $\alpha \in \mathbb{R}^T$ и $\beta \in \mathbb{R}^W$ соответственно:

$$\mathrm{Dir}(\theta_d; \alpha) = \frac{\Gamma(\alpha_0)}{\prod_t \Gamma(\alpha_t)} \prod_t \theta_{td}^{\alpha_t - 1}, \qquad \alpha_t > 0, \quad \alpha_0 = \sum_t \alpha_t, \qquad \theta_{td} > 0, \quad \sum_t \theta_{td} = 1;$$

$$\mathrm{Dir}(\varphi_t; \beta) = \frac{\Gamma(\beta_0)}{\prod_w \Gamma(\beta_w)} \prod_w \varphi_{wt}^{\beta_w - 1}, \quad \beta_w > 0, \quad \beta_0 = \sum_w \beta_w, \quad \varphi_{wt} > 0, \quad \sum_w \varphi_{wt} = 1;$$

где $\Gamma(z)$ — гамма-функция. Гиперпараметры распределения Дирихле связаны с математическим ожиданием порождаемых случайных векторов: $\mathsf{E}\theta_{td} = \frac{\alpha_t}{\alpha_0}$, $\mathsf{E}\varphi_{wt} = \frac{\beta_w}{\beta_0}$.

Введение распределений Дирихле имеет три мотивации.

Во-первых, распределения Дирихле способны порождать как разреженные, так и плотные векторы дискретных распределений, рис. 3. Если вектор параметров состоит из равных значений β_w, то распределение Дирихле называется *симметричным*. При $\beta_w \equiv 1$ симметричное распределение Дирихле совпадает с равномерным распределением на единичном симплексе. Чем меньше β_w, тем ближе к нулю условные вероятности $\varphi_{wt} = p(w\,|\,t)$ в порождаемых векторах φ_t.

Гипотеза о разреженности распределений $\varphi_{wt} = p(w\,|\,t)$ формализует естественное предположение, что каждая тема t имеет *семантическое ядро* — множество термов, характеризующих данную тему и имеющих в ней большие вероятности. Таких термов в каждой теме не может быть много, поскольку большинство термов словаря должны относиться к семантическим ядрам других тем.

Гипотеза о разреженности распределений $\theta_{td} = p(t\,|\,d)$ формализует другое естественное предположение, что каждый документ относится к небольшому числу тем. Трудно представить себе документ обо всех темах (а если это энциклопедия, то в качестве документов стоит взять отдельные её статьи).

Во-вторых, двухуровневая модель порождения данных формализует предположение о существовании тематических кластерных структур в текстовой коллекции. Распределение Дирихле порождает векторы дискретных распределений $\varphi_{wt} = p(w\,|\,t)$,

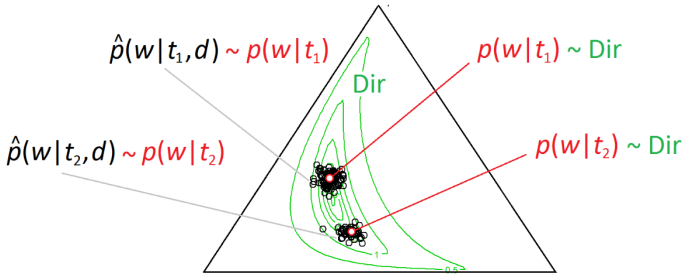


Рис. 4: Распределение $\mathrm{Dir}(\varphi|\alpha)$ порождает векторы тем $\varphi_t = p(w|t)$, которые порождают мультиномиальные распределения $\hat{p}(w|t,d)$ на единичном симплексе в пространстве размерности $|W| = 3$.

которые становятся центрами тематических кластеров. Каждый такой центр порождает тематические части документов — векторы дискретных распределений $p(w|t,d)$, которые плотно группируются вокруг своего центра, рис. 4.

В-третьих, распределение Дирихле является сопряжённым к мультиномиальному распределению. Это чрезвычайно удобно для методов байесовского вывода, которые рассматриваются в следующих главах. Именно математическое удобство предопределило популярность распределения Дирихле и модели LDA в тематическом моделировании, хотя убедительных лингвистических обоснований оно не имеет.

Согласно (22), модели LDA соответствует регуляризатор, с точностью до константы равный сумме логарифмов априорных распределений Дирихле:

$$R(\Phi,\Theta) = \ln \prod_{t\in T} \mathrm{Dir}(\varphi_t;\beta) \prod_{d\in D} \mathrm{Dir}(\theta_d;\alpha) + \mathrm{const} = \\ = \sum_{t\in T}\sum_{w\in W} (\beta_w - 1)\ln\varphi_{wt} + \sum_{d\in D}\sum_{t\in T} (\alpha_t - 1)\ln\theta_{td}. \tag{23}$$

Применение уравнений (17)–(18) к этому регуляризатору даёт формулы М-шага:

$$\varphi_{wt} = \underset{w\in W}{\mathrm{norm}}\big(n_{wt} + \beta_w - 1\big); \qquad \theta_{td} = \underset{t\in T}{\mathrm{norm}}\big(n_{td} + \alpha_t - 1\big).$$

При $\beta_w = 1,\ \alpha_t = 1$ априорное распределение Дирихле совпадает с равномерным распределением на симплексе, формулы М-шага переходят в несмещённые частотные оценки условных вероятностей, а модель LDA переходит в PLSA [69].

При $\beta_w > 1,\ \alpha_t > 1$ регуляризатор имеет сглаживающий эффект, заставляя распределения φ_t и θ_d приближаться к заданным распределениям β и α соответственно.

При $0 < \beta_w < 1,\ 0 < \alpha_t < 1$ распределения φ_t и θ_d, наоборот, отдаляются от распределений β и α. Регуляризатор имеет разреживающий эффект и способен обнулять малые условные вероятности. К сожалению, требование строгой положительности параметров β_w и α_t в распределениях Дирихле ограничивает возможности управления разреженностью матриц Φ, Θ в модели LDA.

Не-байесовская интерпретация и обобщение модели LDA. Регуляризатор (23) можно ввести непосредственно, не прибегая к априорным распределениям Дирихле. Более того, можно снять ограничения положительности с гиперпараметров β_w, α_t и даже задавать их индивидуально для каждой ячейки матриц Φ и Θ.

Введём *обобщённый регуляризатор сглаживания и разреживания*:

$$R(\Phi,\Theta) = \sum_{t\in T}\sum_{w\in W} \beta_{wt}\ln\varphi_{wt} + \sum_{d\in D}\sum_{t\in T} \alpha_{td}\ln\theta_{td}. \tag{24}$$

Подставив этот регуляризатор в (17)–(18), получим формулы М-шага:

$$\varphi_{wt} = \underset{w\in W}{\operatorname{norm}}\big(n_{wt} + \beta_{wt}\big); \tag{25}$$

$$\theta_{td} = \underset{t\in T}{\operatorname{norm}}\big(n_{td} + \alpha_{td}\big). \tag{26}$$

При положительных β_{wt}, α_{td} регуляризатор (24) можно интерпретировать как максимизацию правдоподобия. Он одновременно приближает каждое из распределений φ_t к $\operatorname{norm}_w(\beta_{wt})$ и каждое из распределений θ_d — к $\operatorname{norm}_t(\alpha_{td})$.

При отрицательных β_{wt}, α_{td} минимизация правдоподобия, наоборот, отдаляет столбцы матрицы Φ и Θ от заданных распределений.

Положительное значение гиперпараметра β_{wt} или α_{td} приводит к сглаживанию, отрицательное — к разреживанию соответствующей ячейки матрицы. В отличие от априорных распределений Дирихле, обобщённый регуляризатор не ограничивает параметры снизу, что позволяет свободно управлять разреживанием. Также он снимает ограничение на смешивание отрицательных и положительных значений гиперпараметров в одном столбце.

Разделим положительные и отрицательные параметры в каждом столбце на два отдельных вектора и оба их отнормируем:

$$\beta^+_{wt} = \underset{w\in W}{\operatorname{norm}}\big(\beta_{wt}[\beta_{wt} > 0]\big), \qquad \alpha^+_{td} = \underset{t\in T}{\operatorname{norm}}\big(\alpha_{td}[\alpha_{td} > 0]\big),$$

$$\beta^-_{wt} = \underset{w\in W}{\operatorname{norm}}\big(-\beta_{wt}[\beta_{wt} < 0]\big), \qquad \alpha^-_{td} = \underset{t\in T}{\operatorname{norm}}\big(-\alpha_{td}[\alpha_{td} < 0]\big).$$

Тогда обобщённый регуляризатор (24) можно переписать в виде взвешенной суммы $2|T| + 2|D|$ регуляризаторов, каждый из которых воздействует на отдельный столбец матрицы Φ или Θ:

$$R(\Phi,\Theta) = \sum_{t\in T}\beta^+_{0t}\sum_{w\in W}\beta^+_{wt}\ln\varphi_{wt} + \sum_{d\in D}\alpha^+_{0d}\sum_{t\in T}\alpha^+_{td}\ln\theta_{td} - \\ - \sum_{t\in T}\beta^-_{0t}\sum_{w\in W}\beta^-_{wt}\ln\varphi_{wt} - \sum_{d\in D}\alpha^-_{0d}\sum_{t\in T}\alpha^-_{td}\ln\theta_{td},$$

где β^+_{0t}, α^+_{0d}, β^-_{0t}, α^-_{0d} — положительные нормировочные множители, теперь играющие роль коэффициентов регуляризации.

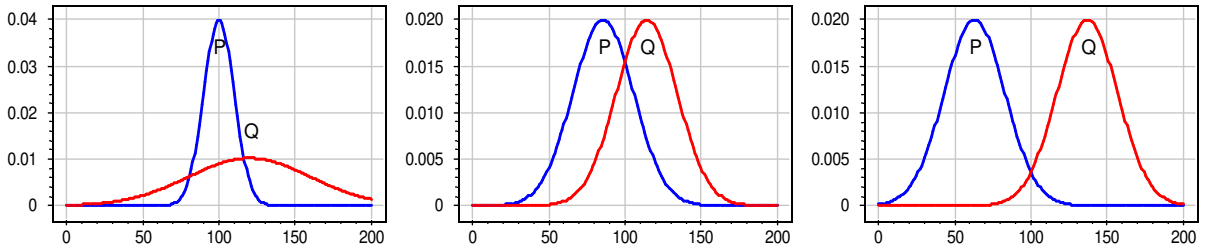
Дивергенция Кульбака–Лейблера (*KL-дивергенция*, относительная энтропия) — это несимметричная функция расстояния между дискретными распределениями $P = (p_i)_{i=1}^n$ и $Q = (q_i)_{i=1}^n$ с совпадающими носителями, $\{i\colon p_i > 0\} = \{i\colon q_i > 0\}$:

$$\mathrm{KL}(P\|Q) = \mathrm{KL}_i(p_i\|q_i) = \sum_{i=1}^n p_i \ln\frac{p_i}{q_i} = H(P,Q) - H(P),$$

где $H(P) = -\sum_i p_i \ln p_i$ — *энтропия* распределения P, $H(P,Q) = -\sum_i p_i \ln q_i$ — *кросс-энтропией* распределений P и Q.

Обозначение KL_i не является общепринятым, но оно удобно для уточнения, по какому индексу производится суммирование.

Перечислим некоторые свойства KL-дивергенции.



$\mathrm{KL}(P\|Q) = 0.44$ $\mathrm{KL}(Q\|P) = 2.97$ | $\mathrm{KL}(P\|Q) = 0.44$ $\mathrm{KL}(Q\|P) = 0.44$ | $\mathrm{KL}(P\|Q) = 2.97$ $\mathrm{KL}(Q\|P) = 2.97$

Рис. 5: Дивергенция Кульбака–Лейблера $\mathrm{KL}(P\|Q)$ является несимметричной мерой вложенности распределения $P = (p_i)_{i=1}^n$ в распределение $Q = (q_i)_{i=1}^n$. Вложенность P в Q численно совпадает на левом и среднем графиках, вложенность Q в P — на левом и правом графиках.

1. $\mathrm{KL}(P\|Q) \geqslant 0$, при этом $\mathrm{KL}(P\|Q) = 0$ тогда и только тогда, когда $P = Q$.

2. KL-дивергенция является мерой вложенности первого распределения во второе: если $\mathrm{KL}(P\|Q) < \mathrm{KL}(Q\|P)$, то P сильнее вложено в Q, чем Q в P, см. рис. 5.

3. Если P — эмпирическое распределение, а $Q(\alpha)$ — параметрическая модель, то минимизация KL-дивергенции эквивалентна минимизации кросс-энтропии $H\big(P, Q(\alpha)\big)$ или максимизации правдоподобия:

$$\mathrm{KL}(P\|Q(\alpha)) = \sum_{i=1}^{n} p_i \ln \frac{p_i}{q_i(\alpha)} \to \min_{\alpha} \quad \Longleftrightarrow \quad \sum_{i=1}^{n} p_i \ln q_i(\alpha) \to \max_{\alpha}.$$

4. Максимизация правдоподобия для тематического моделирования (13) эквивалентна минимизации взвешенной суммы KL-дивергенций по всем документам d между эмпирическими распределениями $\hat{p}(w\,|\,d) = \frac{n_{dw}}{n_d}$ и модельными $p(w\,|\,d)$:

$$\sum_{d\in D} n_d \,\mathrm{KL}_w\Big(\frac{n_{dw}}{n_d} \,\Big\|\, \sum_{t\in T} \varphi_{wt}\theta_{td}\Big) \;\to\; \min_{\Phi,\Theta},$$

где весом документа d является его длина n_d. Если веса n_d убрать, то все документы будут искусственно приведены к одинаковой длине. Такая модификация оптимизационного критерия может быть полезна при моделировании коллекций, содержащих документы одинаковой важности, но существенно разной длины.

5. Максимизация обобщённого регуляризатора сглаживания и разреживания (24) эквивалентна минимизации суммы KL-дивергенций:

$$R(\Phi,\Theta) = -\sum_{t\in T} \beta_{0t}^{+} \,\mathrm{KL}_w\big(\beta_{wt}^{+} \,\big\|\, \varphi_{wt}\big) - \sum_{d\in D} \alpha_{0d}^{+} \,\mathrm{KL}_t\big(\alpha_{td}^{+} \,\big\|\, \theta_{td}\big) + \\ + \sum_{t\in T} \beta_{0t}^{-} \,\mathrm{KL}_w\big(\beta_{wt}^{-} \,\big\|\, \varphi_{wt}\big) + \sum_{d\in D} \alpha_{0d}^{-} \,\mathrm{KL}_t\big(\alpha_{td}^{-} \,\big\|\, \theta_{td}\big),$$

где нормированные векторы гиперпараметров $\beta_t^{\pm}$, $\alpha_d^{\pm}$ в априорных распределениях Дирихле задают желаемое направление сглаживания или разреживания столбцов матриц Φ и Θ, а их нормировочные множители $\beta_{0t}^{\pm}$, $\alpha_{0d}^{\pm}$ играют роль коэффициентов регуляризации.

Выводы по главе

- Тематическая модель LDA — это простой, но устаревший и довольно слабый способ регуляризации. Тем не менее, LDA имеет массу реализаций и чаще всего используется на практике.
- Популярность LDA объясняется не столько его лингвистической обоснованностью, сколько математическим удобством распределения Дирихле в байесовском выводе, см. главу 7.
- Обобщённая модель LDA (без ограничения положительности гиперпараметров) вводится через принцип максимума правдоподобия или минимума KL-дивергенции, без привлечения априорных распределений Дирихле.
- В дальнейшем мы будем активно использовать KL-дивергенцию для конструирования регуляризаторов.

6 Теория EM-алгоритма

В этой главе мы погрузимся в теорию EM-алгоритма. Рассмотрим более общий классический способ его вывода, который позволяет обосновать сходимость. Заодно обогатим его возможностью аддитивной регуляризации. Материал этой главы понадобится только в следующей главе, ещё более теоретической. Если не терпится поскорее узнать о практических аспектах тематического моделирования, то обе главы можно целиком пропустить.

Общий EM-алгоритм с регуляризацией. Исходно EM-алгоритм предназначался для построения широкого класса вероятностных порождающих моделей со скрытыми переменными [57]. Тематическое моделирование является для него лишь частным случаем. Сходимость EM-алгоритма удобно доказывать для общего случая, добавив возможность регуляризации.

Поэтому перейдём к более общим обозначениям:

$X = (d_i, w_i)_{i=1}^n$ — исходные данные, *наблюдаемые переменные*;

$Z = (t_i)_{i=1}^n$ — *скрытые переменные*;

$\Omega = (\Phi, \Theta)$ — параметры порождающей вероятностной модели $p(X, Z \,|\, \Omega)$.

Задача заключается в том, чтобы по выборке X найти параметры модели Ω, при которых достигается максимум *маргинализованного правдоподобия* (marginal likelihood)[1] с регуляризатором $R(\Omega)$:

$$\ln p(X \,|\, \Omega) + R(\Omega) = \ln \sum_Z p(X, Z \,|\, \Omega) + R(\Omega) \;\to\; \max_\Omega. \tag{27}$$

Эта задача неудобна тем, что суммирование под логарифмом производится по всевозможным значениям скрытых переменных Z. В случае тематического моделирования это множество всех n-мерных векторов тем, его мощность равна $|T|^n$.

Теорема 6.1. *Если функционал* (27) *достаточно гладкий, то точка* Ω *его локального максимума удовлетворяет системе уравнений, решение которой методом простых итераций сводится к чередованию двух шагов:*

$$\textit{E-шаг:}\quad q(Z) = p(Z \,|\, X, \Omega); \tag{28}$$

$$\textit{M-шаг:}\quad \sum_Z q(Z) \ln p(X, Z \,|\, \Omega) + R(\Omega) \;\to\; \max_\Omega. \tag{29}$$

Доказательство. Запишем необходимые условия локального экстремума для задачи максимизации гладкого функционала (27):

$$\frac{1}{p(X \,|\, \Omega)} \sum_Z \frac{\partial p(X, Z \,|\, \Omega)}{\partial \Omega} + \frac{\partial R(\Omega)}{\partial \Omega} = 0.$$

[1] Marginal likelihood иногда переводится на русский язык как предельное или неполное правдоподобие.

Из формулы условной вероятности следует $p(X\,|\,\Omega) = \frac{p(X,Z\,|\,\Omega)}{p(Z\,|\,X,\Omega)}$. Воспользуемся этим тождеством, внося $\frac{1}{p(X\,|\,\Omega)}$ под знак суммы:

$$\sum_Z \frac{p(Z\,|\,X,\Omega)}{p(X,Z\,|\,\Omega)} \frac{\partial p(X,Z\,|\,\Omega)}{\partial \Omega} + \frac{\partial R(\Omega)}{\partial \Omega} = 0;$$
$$\sum_Z p(Z\,|\,X,\Omega) \frac{\partial}{\partial \Omega} \ln p(X,Z\,|\,\Omega) + \frac{\partial R(\Omega)}{\partial \Omega} = 0.$$

Полученное уравнение является необходимым условием локального экстремума задачи М-шага (29), если фиксировать распределение $p(Z\,|\,X,\Omega)$ так, чтобы оно не зависело от параметра Ω. Именно это и достигается вычислением $q(Z)$ на Е-шаге (28) и последующей его подстановкой в (29). Таким образом, мы получили систему уравнений, эквивалентную необходимым условиям максимума функционала (27).

Теорема доказана.

Следствие 6.2. *Если задача оптимизации* (27) *имеет ограничения* $g_i(\Omega) \leqslant 0$, $h_j(\Omega) = 0$, *то система уравнений* (28)–(29) *остаётся в силе, при этом задача оптимизации* (29) *имеет те же ограничения.*

Для доказательства необходимые условия локального экстремума заменяются условиями Каруша–Куна–Таккера, в остальном выкладки аналогичны.

Теорема 6.3. *В итерационном процессе* (28)–(29) *значение функционала не уменьшается на каждом шаге.*

Доказательство. Введём произвольное распределение $q(Z)$. Для него справедливо условие нормировки $\sum_Z q(Z) = 1$. Из формулы условной вероятности вытекает тождество $p(X\,|\,\Omega) = \frac{p(X,Z\,|\,\Omega)}{p(Z\,|\,X,\Omega)}$. Следовательно,

$$\ln p(X\,|\,\Omega) = \sum_Z q(Z) \ln p(X\,|\,\Omega) = \sum_Z q(Z) \ln \frac{p(X,Z\,|\,\Omega)}{p(Z\,|\,X,\Omega)}.$$

Добавим и отнимем $\sum_Z q(Z) \ln q(Z)$:

$$\ln p(X\,|\,\Omega) = \underbrace{\sum_Z q(Z) \ln \frac{p(X,Z\,|\,\Omega)}{q(Z)}}_{L(q,\Omega)} + \underbrace{\sum_Z q(Z) \ln \frac{q(Z)}{p(Z\,|\,X,\Omega)}}_{\mathrm{KL}(q(Z)\|p(Z\,|\,X,\Omega)) \geqslant 0}. \tag{30}$$

Второе слагаемое в этой сумме является KL-дивергенцией между двумя распределениями, которая всегда неотрицательна. Поэтому первое слагаемое, обозначенное через $L(q,\Omega)$, является нижней оценкой логарифма правдоподобия $\ln p(X\,|\,\Omega)$.

Для максимизации регуляризованного log-правдоподобия $\ln p(X\,|\,\Omega) + R(\Omega)$ по Ω будем максимизировать его нижнюю оценку поочерёдно то по q, то по Ω:

$$\text{Е-шаг:} \quad L(q,\Omega) + R(\Omega) \;\to\; \max_q;$$
$$\text{М-шаг:} \quad L(q,\Omega) + R(\Omega) \;\to\; \max_\Omega.$$

Максимизация $L(q,\Omega)$ по распределению q эквивалентна минимизации дивергенции $\mathrm{KL}\big(q(Z)\,\big\|\,p(Z\,|\,X,\Omega)\big)$, поскольку их сумма (30) не зависит от q. Минимальное значение дивергенции, равное нулю, достигается при $q(Z)=p(Z\,|\,X,\Omega)$, что совпадает с E-шагом (28). При обнулении дивергенции в (30) нижняя оценка $L(q,\Omega)\leqslant\ln p(X\,|\,\Omega)$ обращается в равенство, а задача максимизации на M-шаге совпадает с (29).

Таким образом, EM-алгоритм (28)–(29) является частным случаем итерационного процесса блочно-покоординатной оптимизации функционала $L(q,\Omega)+R(\Omega)$, на каждом шаге которого значение функционала может только увеличиться.

Теорема доказана.

EM-алгоритм не гарантирует ни достижения максимума с заданной точностью, ни глобальной сходимости. Оптимизационная задача является в общем случае многоэкстремальной. На практике качество решения может зависеть от выбора начального приближения.

Общий EM-алгоритм для ARTM. Применим EM-алгоритм (28)–(29) к задаче тематического моделирования, вернувшись к исходным обозначениям: $X=(d_i,w_i)_{i=1}^n$, $Z=(t_i)_{i=1}^n$, $\Omega=(\Phi,\Theta)$. Согласно следующей теореме, это приведёт к уже знакомой постановке задачи (19), хотя и путём более трудоёмкого доказательства.

Теорема 6.4. *Точка (Φ,Θ) локального максимума регуляризованного логарифма правдоподобия в задаче тематического моделирования* (15) *удовлетворяет системе уравнений, решение которой методом простых итераций сводится к чередованию E- и M-шагов:*

$$\text{E-шаг:}\quad p(t\,|\,d,w)=\underset{t\in T}{\operatorname{norm}}\big(\varphi_{wt}\theta_{td}\big),\quad \text{для всех } d\in D,\ w\in W,\ t\in T; \tag{31}$$

$$\text{M-шаг:}\quad \sum_{d\in D}\sum_{w\in W}\sum_{t\in T} n_{dw}p(t\,|\,d,w)\ln\big(\varphi_{wt}\theta_{td}\big)+R(\Phi,\Theta)\ \to\ \max_{\Phi,\Theta}. \tag{32}$$

Доказательство. Запишем сначала формулу E-шага. Воспользовавшись условием независимости элементов выборки и применив формулу Байеса (5), разложим $q(Z)$ в произведение условных вероятностей тем по всем позициям i:

$$q(Z)=p(Z\,|\,X,\Omega)=\prod_{i=1}^n p(t_i\,|\,d_i,w_i)=\prod_{i=1}^n\underset{t_i}{\operatorname{norm}}\big(\varphi_{w_it_i}\theta_{t_id_i}\big).$$

Подставим в формулу M-шага (29) распределения $p(X,Z\,|\,\Omega)$ и $q(Z)$:

$$\sum_{Z\in T^n} q(Z)\ln p(X,Z\,|\,\Omega)+R(\Omega)\ \to\ \max_{\Omega};$$

$$\sum_{t_1\in T}\cdots\sum_{t_n\in T}\prod_{k=1}^n p(t_k\,|\,d_k,w_k)\ln\prod_{i=1}^n p(d_i,w_i,t_i\,|\,\Omega)+R(\Omega)\ \to\ \max_{\Omega}.$$

Выразим логарифм произведения через сумму логарифмов и переставим местами знаки суммирования:

$$\sum_{i=1}^n\sum_{t_1\in T}\cdots\sum_{t_n\in T}\prod_{k=1}^n p(t_k\,|\,d_k,w_k)\ln p(d_i,w_i,t_i\,|\,\Omega)+R(\Omega)\ \to\ \max_{\Omega}.$$

Заметим, что среди n сумм по $t_i \in T$ нетривиальна всегда только одна — та, для которой $t_i = t_k$. Все остальные суммы расходуются на образование полных вероятностей $\sum_{t_k} p(t_k \,|\, d_k, w_k) = 1$. Таким образом, функция в левой части упрощается:

$$\sum_{i=1}^{n} \sum_{t \in T} p(t \,|\, d_i, w_i) \ln p(d_i, w_i, t \,|\, \Omega) + R(\Omega) \;\to\; \max_{\Omega}.$$

Заменим суммирование по позициям термов i суммированием по документам d, затем по термам w в каждом документе, учитывая каждый терм n_{dw} раз. Затем воспользуемся представлением $p(d, w, t \,|\, \Omega) = p(w \,|\, t, \Phi)\, p(t \,|\, d, \Theta)\, p(d) = \varphi_{wt} \theta_{td} p_d$, отбросив в нём множитель p_d, не зависящий от искомых параметров модели:

$$\sum_{d \in D} \sum_{w \in W} \sum_{t \in T} n_{dw} p(t \,|\, d, w) \ln\bigl(\varphi_{wt} \theta_{td}\bigr) + R(\Phi, \Theta) \;\to\; \max_{\Phi, \Theta}.$$

Теорема доказана.

Таким образом, применение классического EM-алгоритма к задаче тематического моделирования приводит к той же теореме 4.1 через промежуточную форму записи M-шага (32), совпадающую с (19).

Выводы по главе

- Классический EM-алгоритм (28)–(29) решает задачу оценивания параметров вероятностной модели со скрытыми переменными путём максимизации маргинализованного log-правдоподобия. EM-алгоритм для ARTM выводится из него как частный случай.
- EM-алгоритм для тематического моделирования (16)–(18) решает задачу стохастического матричного разложения путём максимизации правдоподобия.
- В тематическом моделировании оба варианта вывода приводят к одному и тому же алгоритму, допускают введение гладких регуляризаторов и располагают доказательствами сходимости.
- В тематическом моделировании вывод EM-алгоритма с помощью леммы о максимизации на единичных симплексах оказывается проще и гибче, чем классический вывод.

7 Байесовское обучение модели LDA

Целью байесовского обучения (Bayesian learning) является получение оценок плотности распределения для параметров вероятностной модели вместо обычных точечных оценок. В анализе данных есть масса практических задач и ситуаций, когда апостериорные оценки плотности действительно необходимы. Однако в практике тематического моделирования всегда используются точечные оценки элементов матриц Φ и Θ, а не плотности распределения $p(\Phi)$, $p(\Theta)$.

Остаётся лишь догадываться, почему байесовское обучение оказалось доминирующим подходом в тематическом моделировании. Возможно, байесовские методы были на возрастающем тренде «кривой Гартнера», когда появилась модель PLSA Томаса Хофманна. Возможно, сказался безусловный научный авторитет авторов модели LDA. Фактически, тематическое моделирование проскочило естественную стадию развития. В регрессионном анализе, обработке сигналов и изображений постановки оптимизационных задач усложнялись постепенно, в том числе путём введения регуляризаторов. Когда возникли приложения, требующие знать больше о распределениях параметров, вместо обычной регуляризации стали применять байесовскую и использовать байесовский вывод. Тематическое моделирование оказалась относительно новой задачей в тот момент, когда байесовское обучение было на пике популярности. Простота и богатые возможности классической не байесовской регуляризации оказались незамеченными. Это упущение как раз и устраняет теория ARTM.

Несмотря на высказанные выше соображения, байесовское обучение крайне важно для тематического моделирования хотя бы потому, что за два десятилетия тысячи публикаций были написаны именно на этом математическом языке.

Мы рассмотрим два наиболее популярных подхода: вариационный байесовский вывод (variational Bayes, VB) [173] и сэмплирование Гиббса (Gibbs sampling, GS) [168]. Для простоты ограничимся моделью LDA и увидим, что оба подхода приводят к EM-подобным алгоритмам, незначительно отличающимся от знакомой нам версии.

Громоздкая техника байесовского вывода, описанная в данной главе, далее использоваться не будет. Байесовские тематические модели, как правило, удаётся переформулировать в терминах регуляризации ещё на этапе постановки задачи, сделав математические выкладки и всё изложение намного яснее и проще. Примеров такой «дебайесизации» моделей будет много в следующих главах.

Концепция байесовского обучения. Пусть X — наблюдаемая выборка данных, $p(X\,|\,\Omega)$ — вероятностная модель данных с параметрами Ω, $p(\Omega\,|\,\gamma)$ — *априорное распределение* в пространстве параметров модели, имеющее гиперпараметры γ. Тогда *апостериорное распределение* параметров, согласно формуле Байеса, имеет вид

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

где символ пропорциональности $\propto$ означает «равно с точностью до нормировки».

Если нам нужна лишь оценка максимума правдоподобия для апостериорного распределения, то достаточно воспользоваться принципом *максимума апостериорной вероятности*, который был рассмотрен в предыдущей главе. Тогда задача сводится к максимизации логарифма правдоподобия с вероятностным регуляризатором:

$$\ln p(\Omega\,|\,X,\gamma) = \ln p(X\,|\,\Omega) + \ln p(\Omega\,|\,\gamma) + \text{const} \;\to\; \max_{\Omega,\gamma}.$$

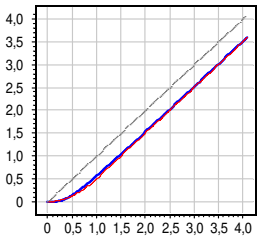
Альтернативный путь, называемый *байесовским выводом*, заключается в том, чтобы вычислить апостериорное распределение $p(\Omega \,|\, X, \gamma)$ в явном виде. Это более сложная задача, поскольку вместо точечной оценки параметра Ω строится его распределение. Зато такой подход даёт гораздо больше информации о параметрах модели.

Свойства распределения Дирихле. Перечислим в справочном порядке некоторые свойства распределения Дирихле, которые понадобятся для вывода модели LDA:

- $\mathsf{E}\theta_t = \int \theta_t \operatorname{Dir}(\theta \,|\, \alpha)\, d\theta = \frac{\alpha_t}{\alpha_0} = \underset{t}{\operatorname{norm}}(\alpha_t)$ — математическое ожидание θ_t;
- $\hat{\theta}_t = \frac{\alpha_t - 1}{\alpha_0 - T} = \underset{t}{\operatorname{norm}}(\alpha_t - 1)$ — мода;
- $\mathsf{D}\theta_t = \frac{\alpha_t(\alpha_0 - \alpha_t)}{\alpha_0^2(\alpha_0 + 1)}$ — дисперсия;
- $\mathsf{E} \ln \theta_t = \int \ln \theta_t \operatorname{Dir}(\theta \,|\, \alpha)\, d\theta = \psi(\alpha_t) - \psi(\alpha_0)$ — математическое ожидание $\ln \theta_t$.

Дигамма-функция $\psi(x) = \frac{\Gamma'(x)}{\Gamma(x)}$ очень похожа на логарифм. Известна простая, но весьма точная аппроксимация экспоненты дигамма-функции (на графике их линии практически неразличимы):

$$E(x) = \exp(\psi(x)) \approx \begin{cases} \frac{x^2}{2}, & 0 \leqslant x \leqslant 1; \\ x - \frac{1}{2}, & 1 \leqslant x. \end{cases}$$



Вариационный байесовский вывод. Идея этого подхода в том, чтобы искать совместное апостериорное распределение для параметров модели и скрытых переменных $p(Z, \Phi, \Theta \,|\, X, \alpha, \beta)$. Непосредственное вычисление данного распределения проблематично. Поэтому находят его приближение в виде разложения на множители, используя *основную теорему вариационного байесовского вывода.*

Теорема 7.1. *Решение задачи* $\mathrm{KL}\big(q(Y) \,\big\|\, p(Y \,|\, X, \gamma)\big) \to \min\limits_{q}$ *в классе распределений* $q(Y) = \prod\limits_{j \in J} q_j(Y_j)$, *факторизованных по блокам переменных* $Y = (Y_j \colon j \in J)$, *удовлетворяет системе уравнений*

$$\ln q_j(Y_j) = \mathsf{E}_{q \setminus j} \ln p(X, Y \,|\, \gamma) + \text{const}, \quad j \in J, \tag{33}$$

где $\mathsf{E}_{q \setminus j}$ — *математическое ожидание по всем переменным* $Y \setminus Y_j$, const — *логарифм нормировочного множителя распределения* q_j.

Доказательство можно найти, например, в [34].

Система уравнений (33) записана в виде, удобном для её численного решения методом простой итерации.

Применим эту теорему к нашему случаю: $Y = (Z, \Phi, \Theta)$, $\gamma = (\alpha, \beta)$. Приблизим распределение $p(Y \,|\, X, \gamma) = p(Z, \Phi, \Theta \,|\, X, \alpha, \beta)$ произведением $n + |T| + |D|$ распределений по блокам переменных t_i, φ_t, θ_d:

$$q(Z, \Phi, \Theta) = \prod_{j \in J} q_j(Z, \Phi, \Theta) = \prod_{i=1}^{n} q_i(t_i) \prod_{t \in T} q_t(\varphi_t) \prod_{d \in D} q_d(\theta_d),$$

где $J = \{1, \dots, n\} \sqcup T \sqcup D$ — индексы всех блоков переменных.

Заметим, что если блоки переменных Y_j независимы, то решение будет не приближённым, а точным. Таким образом, использование вариационного байесовского вывода связано с предположением, что можно пренебречь зависимостями между темами термов t_i, векторами тем φ_t и векторами документов θ_d.

Чтобы записать систему уравнений (33), распишем логарифм распределения $p(X, Z, \Phi, \Theta \,|\, \alpha, \beta)$, переводя слагаемые, не зависящие от переменных t_i, φ_t, θ_d, в const:

$$\begin{aligned}
&\ln p(X, Z, \Phi, \Theta \,|\, \alpha, \beta) = \ln p(X, Z \,|\, \Phi, \Theta)\, p(\Phi \,|\, \beta)\, p(\Theta \,|\, \alpha) = \\
&= \ln \prod_{i=1}^{n} p(d_i, w_i, t_i \,|\, \Phi, \Theta) + \ln \prod_{t \in T} \mathrm{Dir}(\varphi_t \,|\, \beta) + \ln \prod_{d \in D} \mathrm{Dir}(\theta_d \,|\, \alpha) = \\
&= \sum_{i=1}^{n} \ln\bigl(\varphi_{w_i t_i} \theta_{t_i d_i}\bigr) + \sum_{t,w} (\beta_w - 1) \ln \varphi_{wt} + \sum_{d,t} (\alpha_t - 1) \ln \theta_{td} + \mathrm{const}.
\end{aligned}$$

Теперь надо брать математические ожидания $\mathsf{E}_{q \setminus j}$ от этой суммы по всем распределениям $q_t(\varphi_t)$, $q_d(\theta_d)$, $q_i(t_i)$, кроме j-го. Заметим, что если слагаемое S не зависит от j-й переменной, то $\mathsf{E}_{q_j} S = \mathrm{const}$, что сильно упрощает выкладки.

Рассмотрим уравнение (33) относительно $q_t(\varphi_t)$:

$$\begin{aligned}
\ln q_t(\varphi_t) &= \sum_{i=1}^{n} \mathsf{E}_{q_i(t_i)} [t_i = t] \ln \varphi_{w_i t_i} + \sum_{w \in W} (\beta_w - 1) \ln \varphi_{wt} + \mathrm{const} = \\
&= \sum_{i=1}^{n} \sum_{w \in W} [w_i = w] q_i(t) \ln \varphi_{wt} + \sum_{w \in W} (\beta_w - 1) \ln \varphi_{wt} + \mathrm{const} = \\
&= \sum_{w \in W} \Biggl(\underbrace{\sum_{i=1}^{n} [w_i = w] q_i(t)}_{n_{wt}} + \beta_w - 1 \Biggr) \ln \varphi_{wt} + \mathrm{const} = \\
&= \ln \mathrm{Dir}(\varphi_t \,|\, \tilde{\beta}_t).
\end{aligned}$$

Таким образом, $q_t(\varphi_t)$ является распределением Дирихле с параметрами $\tilde{\beta}_{wt} = n_{wt} + \beta_w$, где n_{wt} — оценка числа генераций терма w из темы t. При больших n_{wt} оно сконцентрировано в точке $\varphi_{wt} = \operatorname*{norm}_w(\tilde{\beta}_{wt})$.

Рассмотрим уравнение (33) относительно $q_d(\theta_d)$:

$$\begin{aligned}
\ln q_d(\theta_d) &= \sum_{i=1}^{n} \mathsf{E}_{q_i(t_i)} [d_i = d] \ln \theta_{t_i d_i} + \sum_{t \in T} (\alpha_t - 1) \ln \theta_{td} + \mathrm{const} = \\
&= \sum_{i=1}^{n} [d_i = d] \sum_{t \in T} q_i(t) \ln \theta_{td} + \sum_{t \in T} (\alpha_t - 1) \ln \theta_{td} + \mathrm{const} = \\
&= \sum_{t \in T} \Biggl(\underbrace{\sum_{i=1}^{n} [d_i = d] q_i(t)}_{n_{td}} + \alpha_t - 1 \Biggr) \ln \theta_{td} + \mathrm{const} = \\
&= \ln \mathrm{Dir}(\theta_d \,|\, \tilde{\alpha}_d).
\end{aligned}$$

Таким образом, $q_d(\theta_d)$ является распределением Дирихле с параметрами $\tilde{\alpha}_{td} = n_{td} + \alpha_t$, где n_{td} — оценка числа термов темы t в документе d. При больших n_{td} оно сконцентрировано в точке $\theta_{td} = \underset{t}{\operatorname{norm}}(\tilde{\alpha}_{td})$.

Наконец, рассмотрим уравнение (33) относительно $q_i(t_i)$:

$$\begin{aligned}\ln q_i(t) &= \mathsf{E}_{q\setminus i}\big(\ln \varphi_{w_i t_i} + \ln \theta_{t_i d_i}\big) + \text{const} = \\ &= \mathsf{E}_{q_t(\varphi_t)} \ln \varphi_{w_i t} + \mathsf{E}_{q_d(\theta_d)} \ln \theta_{t_i d} + \text{const}.\end{aligned}$$

Мы уже знаем, что $q_t(\varphi_t)$ и $q_d(\theta_d)$ являются распределениями Дирихле. Воспользуемся известным выражением для математического ожидания логарифма t-й компоненты случайного вектора (θ_t), порождаемого распределением Дирихле:

$$\begin{aligned}\ln q_i(t) = \psi\big(n_{w_i t} + \beta_{w_i}\big) - \psi\big(\textstyle\sum_w (n_{wt} + \beta_w)\big) + \\ + \psi\big(n_{td_i} + \alpha_t\big) - \psi\big(\textstyle\sum_t (n_{td_i} + \alpha_t)\big) + \text{const}.\end{aligned}$$

Логарифмируя и нормируя, получаем распределения переменных t_i:

$$q_i(t) = \underset{t\in T}{\operatorname{norm}} \left(\frac{E\big(n_{w_i t} + \beta_{w_i}\big)}{E\big(\sum_w (n_{wt} + \beta_w)\big)} \cdot \frac{E\big(n_{td_i} + \alpha_t\big)}{E\big(\sum_t (n_{td_i} + \alpha_t)\big)} \right), \tag{34}$$

или, с использованием приближения $E(x) = \exp(\psi(x)) \approx x - \frac{1}{2}$:

$$q_i(t) = \underset{t\in T}{\operatorname{norm}} \left(\frac{n_{w_i t} + \beta_{w_i} - \frac{1}{2}}{n_t + \beta_0 - \frac{1}{2}} \cdot \frac{n_{td_i} + \alpha_t - \frac{1}{2}}{n_{d_i} + \alpha_0 - \frac{1}{2}} \right).$$

Заметим, что формула для $q_i(t)$ похожа на E-шаг (16) в EM-алгоритме для модели LDA: $p(t\,|\,d_i, w_i) = \operatorname{norm}_t(\varphi_{w_i t}\theta_{td_i})$. Более того, аккумулирование счётчиков n_{wt} и n_{td} в точности совпадает с формулами M-шага (17)–(18), если полагать $q_i(t) = p(t\,|\,d_i, w_i)$:

$$n_{wt} = \sum_{i=1}^{n} [w_i = w]\, q_i(t), \qquad n_{td} = \sum_{i=1}^{n} [d_i = d]\, q_i(t).$$

Таким образом, решение системы (33) методом простых итераций приводит к EM-подобному алгоритму. Это немного удивительно, поскольку мы не использовали EM-алгоритм, и даже не решали задачу максимизации правдоподобия.

По окончании итераций искомые параметры модели можно оценить математическим ожиданием апостериорных распределений Дирихле:

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

Основное отличие от EM-алгоритма — в поправках (-1, $-\frac{1}{2}$ или 0) к частотным оценкам условных вероятностей. При $n_{wt}, n_{td} \gg 1$ эти поправки пренебрежимо малы. Они влияют лишь на близкие к нулю условные вероятности φ_{wt} и θ_{td}, которые не являются значимым для тематической модели. Такие отличия в EM-подобных алгоритмах тематического моделирования можно считать несущественными [28].

Сэмплирование Гиббса. Идея этого подхода заключается в том, чтобы сначала оценить скрытые переменные путём сэмплирования $Z \sim p(Z\,|\,X,\alpha,\beta)$, затем найти апостериорное распределение параметров модели $p(\Phi,\Theta\,|\,X,Z,\alpha,\beta)$, при известных X и Z.

Сэмплирование случайного вектора Z из многомерного вероятностного распределения является нетривиальной задачей, но его можно свести к сэмплированию одномерных случайных величин

Теорема 7.2 (о сходимости сэмплирования Гиббса). *Процесс сэмплирования одномерных случайных величин*

$$t_i^{(k+1)} \sim p(t_i\,|\,X, Z_{\setminus i}, \gamma) = \frac{p(X,Z\,|\,\gamma)}{p(X,Z_{\setminus i}\,|\,\gamma)}, \quad i = 1,\ldots,n;$$

где k — номер итерации, $Z_{\setminus i} = (t_1^{(k+1)},\ldots,t_{i-1}^{(k+1)},t_{i+1}^{(k)},\ldots,t_n^{(k)})$, сходится к многомерному распределению $Z \sim p(Z\,|\,X,\gamma)$.

Доказательство можно найти, например, в [34].

Покажем сначала, что если априорное распределение $p(\Phi,\Theta\,|\,\alpha,\beta)$ является произведением распределений Дирихле, то апостериорное распределение принадлежит тому же параметрическому семейству, то есть опять-таки является произведением распределений Дирихле (другими словами, что произведение распределений Дирихле является *сопряжённым* мультиномиальному распределению $p(X,Z\,|\,\Phi,\Theta)$):

$$\begin{aligned}
p(\Phi,\Theta\,|\,X,Z,\alpha,\beta) &\propto p(\Phi,\Theta,X,Z\,|\,\alpha,\beta) \propto p(X,Z\,|\,\Phi,\Theta)\,p(\Phi,\Theta\,|\,\alpha,\beta) \\
&\propto \prod_{d,w,t}(\varphi_{wt}\theta_{td})^{n_{dwt}} \prod_{t\in T}\mathrm{Dir}(\varphi_t\,|\,\beta) \prod_{d\in D}\mathrm{Dir}(\theta_d\,|\,\alpha) \\
&\propto \prod_{t\in T}\prod_{d,w}\varphi_{wt}^{n_{dwt}}\varphi_{wt}^{\beta_w-1} \prod_{d\in D}\prod_{w,t}\theta_{td}^{n_{dwt}}\theta_{td}^{\alpha_t-1} \\
&\propto \prod_{t\in T}\prod_{w}\varphi_{wt}^{n_{wt}+\beta_w-1} \prod_{d\in D}\prod_{t}\theta_{td}^{n_{td}+\alpha_t-1} \\
&\propto \prod_{t\in T}\mathrm{Dir}(\varphi_t\,|\,\tilde\beta_t) \prod_{d\in D}\mathrm{Dir}(\theta_d\,|\,\tilde\alpha_d); \qquad (35)
\end{aligned}$$

где $\tilde\beta_{wt} = n_{wt} + \beta_w$, $\tilde\alpha_{td} = n_{td} + \alpha_t$, счётчики n_{dwt}, n_{wt} и n_{td} определяются согласно (4) через значения скрытых переменных Z.

Чтобы воспользоваться теоремой 7.2, положим $\gamma = (\alpha,\beta)$ и найдём распределение $p(X,Z\,|\,\alpha,\beta)$. Поскольку оно не должно зависеть от параметров (Φ,Θ), по ним придётся взять интеграл. Подынтегральное распределение мы уже вывели в (35), но лишь с точностью до нормировочных множителей. Теперь они нам понадобятся, поэтому разберёмся с ними аккуратнее:

$$\begin{aligned}
p(X,Z\,|\,\alpha,\beta) &= \int_\Phi\int_\Theta p(X,Z\,|\,\Phi,\Theta)p(\Phi,\Theta\,|\,\alpha,\beta)d\Phi d\Theta = \\
&= \int_\Phi\int_\Theta \prod_{w,t}\varphi_{wt}^{n_{wt}} \prod_{t,d}\theta_{td}^{n_{td}} \prod_d p_d^{n_d} \prod_{t\in T}\mathrm{Dir}(\varphi_t\,|\,\beta) \prod_{d\in D}\mathrm{Dir}(\theta_d\,|\,\alpha)d\Phi d\Theta =
\end{aligned}$$

$$= \prod_{t\in T} \frac{\Gamma(\sum_w \beta_w)}{\prod_w \Gamma(\beta_w)} \int_{\varphi_t} \underbrace{\prod_w \varphi_{wt}^{\tilde\beta_{wt}-1} d\varphi_t}_{\propto \mathrm{Dir}(\varphi_t \,|\, \tilde\beta_t)} \prod_{d\in D} p_d^{n_d} \frac{\Gamma(\sum_t \alpha_t)}{\prod_t \Gamma(\alpha_t)} \int_{\theta_d} \underbrace{\prod_t \theta_{td}^{\tilde\alpha_{td}-1} d\theta_d}_{\propto \mathrm{Dir}(\theta_d \,|\, \tilde\alpha_d)} =$$

$$= \prod_{t\in T} \frac{\Gamma(\sum_w \beta_w)}{\prod_w \Gamma(\beta_w)} \frac{\prod_w \Gamma(\tilde\beta_{wt})}{\Gamma(\sum_w \tilde\beta_{wt})} \prod_{d\in D} p_d^{n_d} \frac{\Gamma(\sum_t \alpha_t)}{\prod_t \Gamma(\alpha_t)} \frac{\prod_t \Gamma(\tilde\alpha_{td})}{\Gamma(\sum_t \tilde\alpha_{td})}. \tag{36}$$

Следуя теореме 7.2, необходимо найти распределение $p(X, Z_{\setminus i} \,|\, \alpha, \beta)$, которое отличается от полученного выше $p(X, Z \,|\, \alpha, \beta)$ лишь тем, что оно строится по той же выборке, но без i-го элемента (d_i, w_i, t_i):

$$p(X, Z_{\setminus i} \,|\, \alpha, \beta) = \prod_{t\in T} \frac{\Gamma(\sum_w \beta_w)}{\prod_w \Gamma(\beta_w)} \frac{\prod_w \Gamma(\tilde\beta_{wt} - \delta^i_{wt})}{\Gamma(\sum_w (\tilde\beta_{wt} - \delta^i_{wt}))} \prod_{d\in D} p_d^{n_d} \frac{\Gamma(\sum_t \alpha_t)}{\prod_t \Gamma(\alpha_t)} \frac{\prod_t \Gamma(\tilde\alpha_{td} - \delta^i_{td})}{\Gamma(\sum_t (\tilde\alpha_{td} - \delta^i_{td}))},$$

где $\delta^i_{wt} = [w{=}w_i][t{=}t_i]$, $\delta^i_{td} = [t{=}t_i][d{=}d_i]$. Теперь, согласно формуле из теоремы 7.2, поделим полученные распределения одно на другое, чтобы получить одномерное распределение для сэмплирования темы t_i. При делении в числителе и знаменателе сократятся все множители кроме тех, которые зависят от элемента (d_i, w_i, t_i):

$$p(t_i \,|\, X, Z_{\setminus i}, \alpha, \beta) = \frac{p(X, Z \,|\, \alpha, \beta)}{p(X, Z_{\setminus i} \,|\, \alpha, \beta)} =$$

$$= \frac{\Gamma(n_{w_i t_i} {+} \beta_{w_i})\, \Gamma(\sum_w (n_{wt_i} {+} \beta_w) {-} 1)}{\Gamma(n_{w_i t_i} {+} \beta_{w_i} {-} 1)\, \Gamma(\sum_w (n_{wt_i} {+} \beta_w))} \frac{\Gamma(n_{t_i d_i} {+} \alpha_{t_i})\, \Gamma(\sum_t (n_{td_i} {+} \alpha_t) {-} 1)}{\Gamma(n_{t_i d_i} {+} \alpha_{t_i} {-} 1)\, \Gamma(\sum_t (n_{td_i} {+} \alpha_t))}.$$

Для упрощения воспользуемся свойством гамма-функции $\frac{\Gamma(x)}{\Gamma(x-1)} = x - 1$:

$$p(t \,|\, X, Z_{\setminus i}, \alpha, \beta) = \operatorname*{norm}_{t\in T} \left(\frac{n_{w_i t} + \beta_{w_i} - 1}{\sum_w (n_{wt} + \beta_w) - 1} \cdot \frac{n_{td_i} + \alpha_t - 1}{\sum_t (n_{td_i} + \alpha_t) - 1} \right).$$

Данное выражение похоже на частотную оценку условной вероятности на E-шаге EM-алгоритма для моделей PLSA и LDA: $p(t \,|\, d_i, w_i) = \mathrm{norm}_t(\varphi_{w_i t} \theta_{t d_i})$. Таким образом, мы снова получаем EM-подобный алгоритм, хотя задача максимизации правдоподобия даже не ставилась. Как и в случае с вариационными байесовским выводом, алгоритм отличается некоторой смещённостью частотных оценок условных вероятностей. Главное его отличие в том, что для каждого (d_i, w_i), $i = 1, \ldots, n$, происходит сэмплирование только одной темы t_i, которая и участвует в аккумулировании счётчиков n_{wt} и n_{td}:

$$n_{wt} = \sum_{i=1}^{n} [w_i{=}w]\, [t_i{=}t], \qquad n_{td} = \sum_{i=1}^{n} [d_i{=}d]\, [t_i{=}t].$$

Фактически, на M-шаге суммируются не сами распределения $p(t \,|\, d_i, w_i)$, а их эмпирические оценки $p_i(t) = [t{=}t_i]$ по единственной сэмплированной теме t_i. Сумма таких оценок сходится к сумме исходных распределений, согласно закону больших чисел.

По окончании итераций искомые параметры модели можно оценить через математическое ожидание апостериорных распределений Дирихле из (35):

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

Алгоритм 4. Сэмплирование Гиббса.

Вход: коллекция D, число тем $|T|$, параметры α, β;
Выход: распределения Φ и Θ;

$n_{wt}, n_{td}, n_t, n_d := 0$ для всех $d \in D,\ w \in W,\ t \in T$;
для всех итераций $k := 1, \ldots, k_{\max}$
  для всех $i = 1, \ldots, n$ взять документ $d := d_i$, терм $w := w_i$
    если $k \geqslant 2$ **то** $t := t_i$; $--n_{wt}$; $--n_{td}$; $--n_t$; $--n_d$;
    $p(t\,|\,d, w) = \underset{t \in T}{\operatorname{norm}}\left(\frac{n_{wt}+\beta_w}{n_t+\beta_0} \cdot \frac{n_{td}+\alpha_t}{n_d+\alpha_0}\right)$ для всех $t \in T$;
    сэмплировать одну тему t из распределения $p(t\,|\,d, w)$;
    $t_i := t$; $++n_{wt}$; $++n_{td}$; $++n_t$; $++n_d$;
$\varphi_{wt} := n_{wt}/n_t$ для всех $w \in W,\ t \in T$;
$\theta_{td} := n_{td}/n_d$ для всех $d \in D,\ t \in T$;

В алгоритме 4 показана идея реализации, предложенная в [168]. Каждая итерация k в процессе сэмплирования соответствует одному проходу коллекции. Когда для позиции i сэмплируется тема t_i, всем предыдущим позициям $1, \ldots, i-1$ уже присвоены темы на данной итерации, а все последующие позиции $i+1, \ldots, n$ сохраняют темы с предыдущей $(k-1)$-й итерации — в точности как того требует теорема 7.2. Сэмплированная тема запоминается в переменной t_i, и на следующей итерации, когда эта тема изменится, счётчики n_{wt} и n_{td} будут уменьшены на единицу для старой темы, и увеличены на единицу для новой.

Значительное конструктивное сходство EM-подобных алгоритмов PLSA, MAP, VB, GS и ещё нескольких их вариантов было впервые отмечено в [28].

Сэмплирование единственной темы на E-шаге $t_i \sim p(t\,|\,d, w)$ можно рассматривать как отдельную эвристику, совместимую с любым EM-подобным алгоритмом тематического моделирования [7]. Эксперименты показали, что сэмплирование несущественно влияет на сходимость и другие свойства модели. Как эвристика, оно может свободно сочетаться с любыми регуляризаторами.

Оптимизация гиперпараметров в модели LDA. Во многих работах, начиная с [168], используются симметричные априорные распределения Дирихле с параметрами $\alpha_t = 50/|T|$, $\beta_w = 0.01$. Более тонкие исследования [186] показали, что лучше оптимизировать вектор $\alpha = (\alpha_1, \ldots, \alpha_T)$ в несимметричном распределении $\mathrm{Dir}(\theta_d\,|\,\alpha)$ и подбирать скалярный числовой параметр $\beta_w \ll 1$ в симметричном распределении $\mathrm{Dir}(\varphi_t\,|\,\beta)$.

Для оптимизации вектора α в [186] предлагается максимизировать правдоподобие $P(X, Z\,|\,\alpha, \beta)$ при фиксированных Z и β. Отбрасывая в (36) множители, не зависящие от α, получаем задачу оптимизации:

$$P(X\,|\,\alpha) = \prod_{d \in D} \frac{\Gamma(\alpha_0)}{\Gamma(n_d + \alpha_0)} \prod_{t \in T} \frac{\Gamma(n_{td} + \alpha_t)}{\Gamma(\alpha_t)} \to \max_{\alpha}.$$

В диссертации [185] сравнивалось более десятка численных методов её решения. Самый простой из них — метод неподвижной точки [127]. В нём вектор $\alpha = (\alpha_t)$

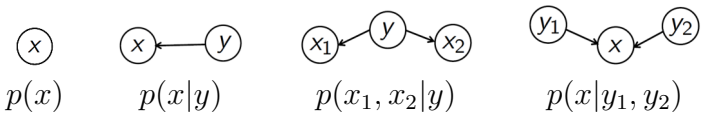


Рис. 6: Отображение условных зависимостей в графической нотации.

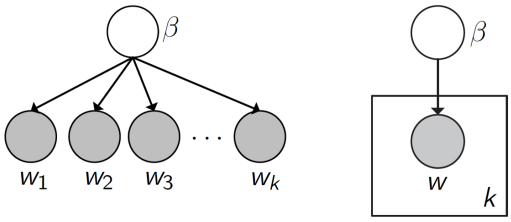


Рис. 7: Графическая нотация выборки $w_1, \ldots, w_k$, порождаемой распределением $\beta_w = p(w)$.

пересчитывается по рекуррентной формуле

$$\alpha_t := \alpha_t \frac{\sum_d \psi(n_{td} + \alpha_t) - \psi(\alpha_t)}{\sum_d \psi(n_d + \alpha_0) - \psi(\alpha_0)},$$

где $\psi(x)$ — дигамма-функция. Эта формула встраивается в итерационный процесс EM-алгоритма между проходами по коллекции.

Эксперименты в [186] показали, что оптимизация вектора α повышает правдоподобие и скорость сходимости EM-алгоритма. В случае сильной *несбалансированности тем* (когда в коллекции одних тем больше, чем других, в разы или даже на порядки) оптимизация вектора α приводит к более естественному неравномерному распределению коллекции по темам.

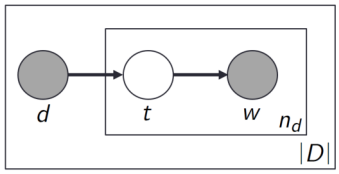
Графическая нотация. Вероятностные тематические модели PLSA и LDA являются частными случаями графовых моделей. В *графовой вероятностной модели* (probabilistic graphical model, PGM) зависимости между случайными величинами представляются в виде графа. Вершины графа соответствуют случайным переменным, рёбра — непосредственным вероятностным взаимосвязям между ними. На рис. 6 показаны примеры зависимостей между случайными переменными. Наблюдаемые переменные изображаются закрашенным кружком. Выборка переменных, генерируемых одним распределением, изображается прямоугольником, рис. 7. Такие изображения принято называть *графической нотацией* (plate notation).

Альтернативной текстуальной формой представления вероятностной модели является *генеративный сценарий* (generative story), который описывает алгоритм порождения данных. На рис. 8 показаны оба представления, графическое и текстуальное, для тематических моделей PLSA и LDA.

Подборка графических нотаций на рис. 9 иллюстрирует большое структурное разнообразие вероятностных тематических моделей. Хотя наглядность является неоспоримым преимуществом, недостатков у графической нотации гораздо больше. Главный — неполнота и неоднозначность интерпретации. Некоторые аспекты моделирования не могут быть отображены общепринятыми средствами графической нотации, и авторы вынуждены изобретать собственные приёмы визуализации. В результате модель может восстанавливаться по картинке неоднозначно. Генеративный сценарий лишен этого недостатка и даёт более точное описание модели.

Модель PLSA:

каждый $d \in D$ порождает скрытые темы:
$t_i \sim p(t \mid d),\ \ i = 1, \ldots, n_d$;
каждая тема t_i порождает терм:
$w_i \sim p(w \mid t_i),\ \ i = 1, \ldots, n_d$.



Модель LDA:

α порождает векторы документов:
$\theta_d \sim \mathrm{Dir}(\theta \mid \alpha),\ \ d \in D$;
β порождает векторы тем:
$\varphi_t \sim \mathrm{Dir}(\varphi \mid \beta),\ \ t \in T$;
далее как в PLSA.

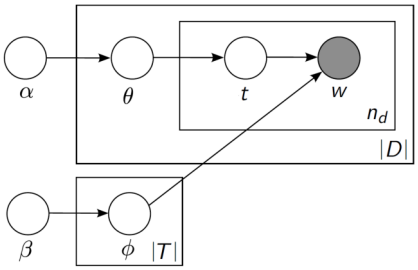


Рис. 8: Генеративная история и графическая нотация для моделей PLSA и LDA.

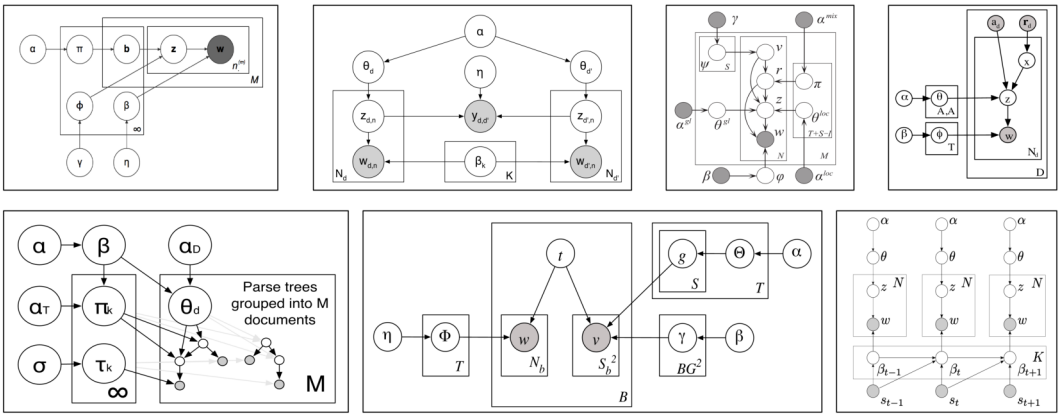


Рис. 9: Примеры графических нотаций из публикаций по тематическому моделированию.

Графическая нотация и генеративный сценарий полезны для понимания модели. Однако оба они описывают лишь прямую задачу — процесс генерации данных по известным параметрам модели. Тогда как в тематическом моделировании решается обратная задача — требуется восстановить значения параметров по наблюдаемым данным. Обоих визуальных представлений не достаточно, чтобы однозначно воспроизвести переход от порождающей модели к алгоритму её обучения. К сожалению, во многих публикациях этот переход опускается для краткости изложения, что отнюдь не добавляет им ясности.

Сравнение ARTM и байесовского подхода. В практике тематического моделирования байесовский вывод апостериорного распределения $p(\Omega \mid X, \gamma)$ производится исключительно ради получения точечной оценки параметров Ω:

$$\mathrm{Posterior}(\Omega \mid X, \gamma) \ \propto\ p(X \mid \Omega)\, \mathrm{Prior}(\Omega \mid \gamma);$$
$$\Omega := \arg\max_{\Omega} \mathrm{Posterior}(\Omega \mid X, \gamma).$$

Максимизация апостериорной вероятности (maximum a posteriori probability, MAP) даёт точечную оценку Ω, минуя промежуточный шаг приближённого и тру-

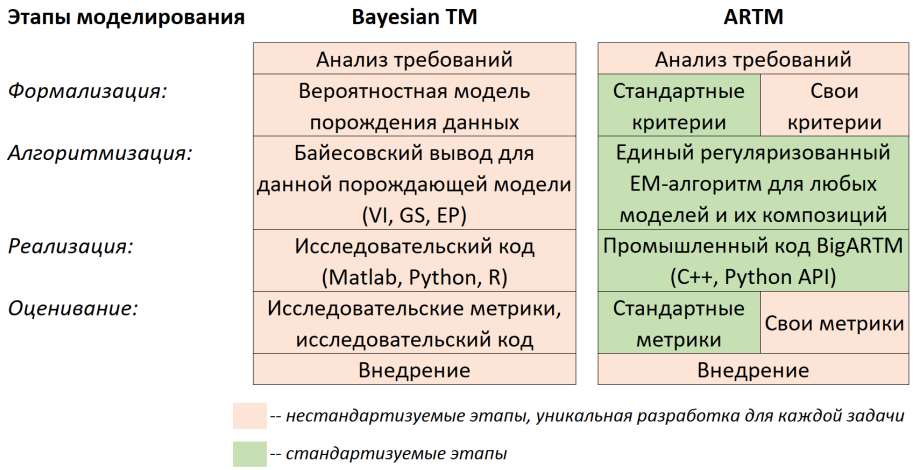


Рис. 10: В отличие от байесовского обучения, ARTM стандартизирует реализацию тематических моделей в виде модульной расширяемой библиотеки регуляризаторов. Процедуры хранения и передачи данных, распараллеливания EM-алгоритма, оценивания качества моделей являются общими для широкого класса моделей и их композиций.

доёмкого вывода Posterior:

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

Многокритериальная аддитивная регуляризация (ARTM) обобщает MAP на любые регуляризаторы, в том числе не имеющие вероятностной интерпретации, а также их произвольные комбинации:

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

К недостаткам байесовского вывода можно отнести техническую сложность формализации требований к модели посредством оптимизационных критериев. Единственным механизмом учёта требований является априорное распределение. Однако байесовский вывод в случае $\mathrm{Prior}(\Omega) \propto \prod_i \exp(\tau_i R_i(\Omega))$, является трудной математической задачей. Когда априорные распределения не являются распределениями Дирихле, байесовский вывод заметно усложняется.

В байесовском тематическом моделировании распределение Дирихле оказывается «на особом положении». 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 оснований предпочитать распределение Дирихле другим регуляризаторам.

Постановка оптимизационной задачи MAP позволяет интерпретировать логарифм априорного распределения как вероятностный регуляризатор, отделить его от конкретной модели и использовать в любых других моделях.

Аддитивность регуляризаторов приводит к модульной технологии тематического моделирования. Каждый регуляризатор R_i реализуется в виде программного модуля, который добавляет к счётчикам n_{wt} и n_{td} в формулах M-шага регуляризационные поправки $\tau_i \varphi_{wt} \frac{\partial R}{\partial \varphi_{wt}}$ и $\tau_i \theta_{td} \frac{\partial R}{\partial \theta_{td}}$ соответственно. Таких модулей можно подключить сколько угодно на этапе построения модели. При этом коэффициенты регуляризации τ_i

можно менять в ходе итерационного процесса по какой угодно стратегии, управляя совокупным воздействием регуляризаторов на модель.

Эти возможности поддерживаются библиотекой `BigARTM` [9, 179]. Комбинирование готовых регуляризаторов при решении прикладных задач позволяет строить модели с заданными свойствами без дополнительных математических выкладок и программирования, рис. 10.

Создание модульной технологии комбинирования моделей в рамках байесовского подхода сильно затруднено, и, насколько нам известно, такие попытки даже не предпринимались.

Выводы по главе

- VB и GS являются наиболее часто используемыми техниками байесовского вывода в тематическом моделировании.
- В дальнейшем мы будем рассматривать все тематические модели в терминах классической регуляризации, даже если в исходных работах они строились байесовскими методами.
- Реформализация байесовских моделей через классическую регуляризацию, как правило, приводит к упрощению изложения при несущественных модификациях в EM-подобных алгоритмах.
- Далее графическая нотация не используется, поскольку в ARTM для формализации и понимания моделей нет необходимости в наглядной визуализации порождающего процесса.

8 Разреживание, сглаживание, декоррелирование

Отказ от априорных распределений Дирихле позволяет обобщить модель LDA: снять ограничения на знаки гиперпараметров в (23) и свободнее обращаться со сглаживанием и разреживанием для улучшения интерпретируемости тем.

Гипотеза разреженности: каждая тема характеризуется небольшим числом термов; каждый документ относится к небольшому числу тем; значительная часть условных вероятностей φ_{wt} и θ_{td} равны нулю.

Разреженность представляется естественным необходимым условием интерпретируемости тематической модели. Кроме того, использование разреженных матриц может сокращать время вычислений и расход памяти как в процессе построения модели, так и при дальнейшем её использовании.

Многократные попытки разреживания модели LDA [162, 62, 189, 99, 48] приводили к её переусложнению из-за внутреннего противоречия между разреженностью и положительностью параметров в распределении Дирихле. Проблема решается при использовании обобщённого регуляризатора сглаживания и разреживания (24), в котором гиперпараметры β_{wt}, α_{td} могут принимать отрицательные значения:

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

Такой способ разреживания был впервые предложен, по всей видимости, в динамической модели PLSA для обработки видеопотоков [175], где документами являлись короткие видеофрагменты, термами — признаки на изображениях, темами — появление определённого объекта в течение определённого времени (например, проезд автомобиля через перекрёсток). Разреживать распределения сильнее, чем это делает LDA, потребовались для описания тем с коротким «временем жизни».

Частичное обучение. В процессе создания, оценивания или использования тематической модели эксперты, асессоры или пользователи могут отмечать в темах релевантные или нерелевантные термы и документы. Размеченные данные позволяют фиксировать интерпретации тем и повышают устойчивость модели. Разметка может затрагивать лишь часть документов и тем, поэтому её использование относится к задачам *частичного обучения* (semi-supervised learning).

Пусть для каждой темы $t \in T$ заданы четыре подмножества:
W_t^+ — «белый список» релевантных термов;
W_t^- — «чёрный список» нерелевантных термов;
D_t^+ — «белый список» релевантных документов;
D_t^- — «чёрный список» нерелевантных документов.

Частичное обучение по релевантности является частным случаем регуляризатора сглаживания и разреживания при

$$\beta_{wt} = \beta_+[w \in W_t^+] - \beta_-[w \in W_t^-],$$
$$\alpha_{td} = \alpha_+[d \in D_t^+] - \alpha_-[d \in D_t^-],$$

где $\beta_\pm$ и $\alpha_\pm$ — коэффициенты регуляризации.

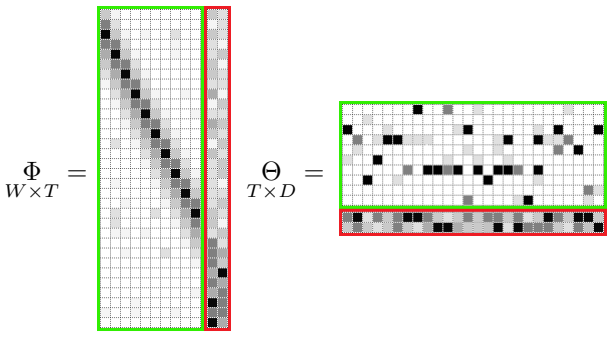


Рис. 11: Структура разреженности матриц Φ и Θ с предметными и фоновыми темами.

Предметные и фоновые темы. Чтобы модель была интерпретируемой, каждая тема должна иметь *семантическое ядро* — множество термов, характеризующих определённую предметную область и редко употребляемых в других темах. Для этого матрицы Φ и Θ должны иметь структуру разреженности, аналогичную показанной на рис. 11. Множество тем разбивается на два подмножества, $T = S \sqcup B$.

Предметные темы $t \in S$ содержат термины предметных областей. Их распределения $p(w\,|\,t)$ разрежены и существенно различны (декоррелированы). Распределения $p(t\,|\,d)$ также разрежены в предметных темах $t \in S$, если предполагать, что каждый документ содержит лишь небольшую долю предметных тем.

Фоновые темы $t \in B$ содержат слова общей лексики, которых не должно быть в предметных темах. Их распределения $p(w\,|\,t)$, возможно, менее разрежены по сравнению с предметными темами и существенно отличаются от них. Распределения $p(t\,|\,d)$ должны быть сильно сглажены в фоновых темах $t \in S$, так как слова общей лексики составляют значительную долю даже в узко специализированных документах. Тематическую модель с несколькими фоновыми темами можно рассматривать как обобщение робастных моделей [46, 147], в которых использовалось только одно фоновое распределение.

Сфокусированный тематический поиск. Частичное обучение тем можно рассматривать как разновидность тематического информационного поиска. В качестве запроса задаётся *семантическое ядро* одной или нескольких тем. Это может быть любой фрагмент текста, «белый список» термов (seed words) или *z-метки* — темы, приписанные отдельным словам или фрагментам в документах [24]. Тематическая поисковая система должна не только найти и ранжировать релевантные документы, но и разложить поисковую выдачу по темам. В типичных приложениях релевантный контент составляет ничтожно малую долю коллекции. Тем не менее, именно этот контент должен быть тщательно систематизирован. Образно говоря, требуется «искать и классифицировать иголки в стоге сена» [42]. Темы становятся элементом графического интерфейса пользователя, инструментом навигации и понимания текстовой коллекции. Отсюда важность требования интерпретируемости каждой темы.

Частичное обучение использовалось для поиска и кластеризации новостей [84], поиска в социальных медиа информации, связанной с болезнями, симптомами и методами лечения [140, 141], с преступностью и экстремизмом [110, 161], с национальностями и межнациональными отношениями [42, 93, 137].

В модели ATAM (ailment topic aspects model) в качестве сглаживающего распределения β_{wt} использовалась большая коллекция медицинских статей [141].

В моделях SSLDA (semi-supervised LDA) и ISLDA (interval semi-supervised LDA) для поиска этно-релевантных тем в постах социальных сетей использовалось сглаживание по словарю из нескольких сотен этнонимов [42]. В модели SSLDA для каждой этно-релевантной темы задаётся свой словарь этнонимов, связанных с одним определённым этносом. В модели ISLDA множество тем разбивается на интервалы, и для всех тем каждого интервала задаётся общий словарь этнонимов. Преимущество этих моделей в том, что интерпретация каждой темы известна заранее. Недостатки в том, что трудно предугадывать число тем для каждой этничности и строить полиэтничные темы для выявления межэтнических отношений.

Для решения этих проблем в [25, 26] был предложен подход на основе ARTM. Задавалось общее число этно-релевантных тем и к ним, как к столбцам матрицы Φ, применялось сглаживание по словарю этнонимов. Тематическая модель сама определяла, как эти темы разделятся по этничностям. Альтернативный подход заключается в том, чтобы придать этнонимам больший вес, сделав их отдельной модальностью (понятие модальности вводится в главе 9).

Любые подходы к сфокусированному тематическому поиску требуют ручной проверки и интерпретации найденных тем.

Декоррелирование тем. Тематическая модель не должна содержать дублирующих или похожих тем. Чем различнее темы, тем информативнее модель. Если темы разделяются на предметные и фоновые, то все они должны существенно отличаться друг от друга.

Для повышения различности тем будем минимизировать сумму попарных скалярных произведений $\langle \varphi_t, \varphi_s \rangle = \sum_w \varphi_{wt}\varphi_{ws}$ между столбцами матрицы Φ. Получим регуляризатор:

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

Формула M-шага, согласно (17), имеет вид

$$\varphi_{wt} = \underset{w \in W}{\operatorname{norm}} \Big(n_{wt} - \tau \varphi_{wt} \sum_{s \in T \setminus t} \varphi_{ws} \Big). \tag{37}$$

Этот регуляризатор контрастирует строки матрицы Φ. В каждой строке, независимо от остальных, вероятности φ_{wt} наиболее значимых тем терма w увеличиваются, вероятности остальных тем уменьшаются и могут обращаться в нуль. Разреживание — это сопутствующий эффект декоррелирования. В [171] был замечен ещё один полезный эффект: слова общей лексики группируются в отдельные темы. Эксперименты с комбинированием регуляризаторов сглаживания, разреживания и декоррелирования в ARTM подтверждают это наблюдение [8, 181, 180].

Декоррелирование впервые было предложено в модели TWC-LDA (topic-weak-correlated LDA) в рамках байесовского подхода [171]. Соответствующее априорное распределение не является сопряжённым к мультиномиальному, поэтому байесовский вывод сталкивается с техническими трудностями. В ARTM расчётная формула M-шага (37) выводится в одну строку.

Комбинирование регуляризаторов сглаживания фоновых тем, разреживания предметных тем в матрице Θ и декоррелирования столбцов матрицы Φ использовалось во многих работах для улучшения интерпретируемости тем [8, 180, 181, 182, 20]. Подобрав коэффициенты регуляризации, можно одновременно значительно улучшить несколько критериев качества тем (разреженность, контрастность, чистоту и когерентность) при незначительной потере правдоподобия [181]. По результатам экспериментов в [181] были выработаны основные рекомендации: декоррелирование и сглаживание включать сразу, разреживание — после 10–20 итераций, когда образуется тенденция к сходимости параметров модели.

Та же комбинация регуляризаторов существенно улучшала качество тематического разведочного поиска в [20, 80, 81], хотя никакие критерии качества поиска непосредственно не оптимизировались.

Выводы по главе

- Разреживание, сглаживание и декоррелирование является минимальным «джентльменским набором» регуляризаторов, применимым к большинству прикладных задач.
- Обобщённая модель LDA, в которой снято ограничение положительности гиперпараметров, находит множество применений: разреживание, частичное обучение, выделение фоновых тем, сфокусированный тематический поиск.

9 Моделирование мультимодальных данных

Мультимодальная тематическая модель описывает документы, содержащие метаданные наряду с основным текстом. Метаданные помогают более точно определять тематику документов, и, наоборот, тематическая модель может использоваться для выявления семантики метаданных или предсказания пропущенных метаданных.

Каждый тип метаданных образует отдельную *модальность* со своим словарём. Слова естественного языка, словосочетания [188, 197], теги [95], именованные сущности [131] — это примеры текстовых модальностей. В мультиязычных тематических моделях параллельных текстов модальностями являются языки [183]. Для анализа коротких текстов с опечатками используют модальность буквенных n-грамм, что позволяет улучшать качество информационного поиска [79].

Примерами нетекстовых модальностей являются (рис. 12): авторы [155], моменты времени [174, 208, 175], классы, жанры или категории [156, 212], цитируемые или цитирующие документы [59] или авторы [88], пользователи электронных библиотек, социальных сетей или рекомендательных систем [100, 165, 190, 204, 205], графические элементы изображений [39, 78, 105], рекламные объявления на веб-страницах [144].

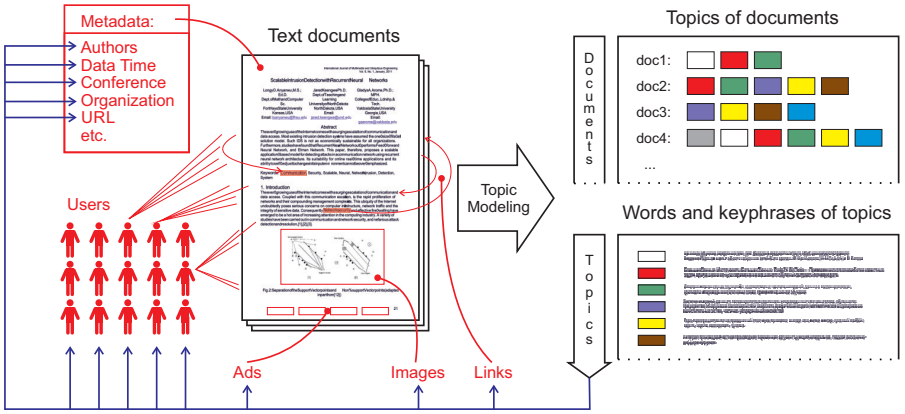


Рис. 12: Обычная тематическая модель определяет распределение тем в каждом документе $p(t|d)$ и распределение термов в каждой теме $p(w|t)$. Мультимодальная модель распространяет семантику тем на элементы всех остальных модальностей, в том числе нетекстовые.

Мультимодальная ARTM. Все перечисленные выше случаи, несмотря на разнообразие интерпретаций, описываются в ARTM единым формализмом модальностей. Каждый документ рассматривается как универсальный контейнер, содержащий термы различных модальностей, включая обычные слова.

Пусть M — множество модальностей. Каждая модальность имеет свой словарь термов W^m, $m \in M$. Эти множества попарно не пересекаются. Их объединение будем обозначать через W. Модальность терма $w \in W$ будем обозначать через $m(w)$.

Тематическая модель модальности m аналогична модели (2):

$$p(w|d) = \sum_{t\in T} p(w|t)\,p(t|d) = \sum_{t\in T} \varphi_{wt}\theta_{td}, \qquad w \in W^m, \quad d \in D. \tag{38}$$

Каждой модальности m соответствует стохастическая матрица $\Phi_m = \big(\varphi_{wt}\big)_{W^m\times T}$. Совокупность матриц Φ_m, если их записать в столбец, образует $W\times T$-матрицу Φ. Распределение тем в каждом документе является общим для всех модальностей.

Мультимодальная модель строится путём максимизации взвешенной суммы логарифмов правдоподобия модальностей и регуляризаторов. Веса τ_m позволяют сбалансировать модальности по их важности и с учётом их частотности в документах:

$$\sum_{m\in M}\tau_m\sum_{d\in D}\sum_{w\in W^m} n_{dw}\ln\sum_{t\in T}\varphi_{wt}\theta_{td} + R(\Phi,\Theta) \;\to\; \max_{\Phi,\Theta}; \tag{39}$$

$$\sum_{w\in W^m}\varphi_{wt}=1;\quad \varphi_{wt}\geqslant 0; \qquad \sum_{t\in T}\theta_{td}=1;\quad \theta_{td}\geqslant 0. \tag{40}$$

Теорема 9.1. *Пусть функция $R(\Phi,\Theta)$ непрерывно дифференцируема. Точка (Φ,Θ) локального экстремума задачи (39)–(40) удовлетворяет системе уравнений со вспомогательными переменными $p_{tdw}=p(t\,|\,d,w)$, если из решения исключить нулевые столбцы матриц Φ_m, Θ:*

$$p_{tdw} = \underset{t\in T}{\operatorname{norm}}\big(\varphi_{wt}\theta_{td}\big); \tag{41}$$

$$\varphi_{wt} = \underset{w\in W^m}{\operatorname{norm}}\left(n_{wt}+\varphi_{wt}\frac{\partial R}{\partial\varphi_{wt}}\right); \qquad n_{wt}=\sum_{d\in D}\tau_{m(w)}n_{dw}p_{tdw}; \tag{42}$$

$$\theta_{td} = \underset{t\in T}{\operatorname{norm}}\left(n_{td}+\theta_{td}\frac{\partial R}{\partial\theta_{td}}\right); \qquad n_{td}=\sum_{w\in W}\tau_{m(w)}n_{dw}p_{tdw}. \tag{43}$$

Доказательство аналогично теореме 4.1, которая является частным случаем теоремы 9.1 для случая одной модальности, $|M|=1$, $\tau_m=1$. Переход от одной модальности к произвольному числу модальностей сводится к двум поправкам:

1) исходные данные n_{dw} домножаются на веса модальностей $\tau_{m(w)}$;

2) матрица Φ разбивается на блоки Φ_m, которые нормируются по-отдельности.

В проекте `BigARTM` реализована возможность комбинировать любое число модальностей с любыми регуляризаторами [25].

Мультиязычные тематические модели используются для кросс-язычного информационного поиска, когда по запросу на одном языке требуется найти схожие документы на другом языке. Для связывания языков используются параллельные тексты или двуязычные словари. Первые мультиязычные тематические модели появились почти одновременно [56, 124, 136] и представляли собой мультимодальную модель, в которой модальностями являются языки, и каждая связка параллельных текстов объединяется в один документ. Оказалось, что связывания документов достаточно для синхронизации тем в разных языках и кросс-язычного поиска. Попытки более точного и трудоёмкого выравнивания по предложениям или по словам практически не улучшают качество поиска. Обстоятельный обзор мультиязычных тематических моделей можно найти в [183].

На рис. 13 показаны некоторые из 400 тем, построенных по 216 175 парам русских и английских статей Википедии [179]. Для связывания языков использовались только модальности, выравнивания и словари не использовались. Асессор оценил 396 тем из 400 как хорошо интерпретируемые.

Для использования двуязычного словаря в [10] был предложен регуляризатор сглаживания. Он формализует предположение, что если слово u в языке k является

Тема №68				Тема №79			
research	4.56	институт	6.03	goals	4.48	матч	6.02
technology	3.14	университет	3.35	league	3.99	игрок	5.56
engineering	2.63	программа	3.17	club	3.76	сборная	4.51
institute	2.37	учебный	2.75	season	3.49	фк	3.25
science	1.97	технический	2.70	scored	2.72	против	3.20
program	1.60	технология	2.30	cup	2.57	клуб	3.14
education	1.44	научный	1.76	goal	2.48	футболист	2.67
campus	1.43	исследование	1.67	apps	1.74	гол	2.65
management	1.38	наука	1.64	debut	1.69	забивать	2.53
programs	1.36	образование	1.47	match	1.67	команда	2.14
Тема №88				**Тема №251**			
opera	7.36	опера	7.82	windows	8.00	windows	6.05
conductor	1.69	оперный	3.13	microsoft	4.03	microsoft	3.76
orchestra	1.14	дирижер	2.82	server	2.93	версия	1.86
wagner	0.97	певец	1.65	software	1.38	приложение	1.86
soprano	0.78	певица	1.51	user	1.03	сервер	1.63
performance	0.78	театр	1.14	security	0.92	server	1.54
mozart	0.74	партия	1.05	mitchell	0.82	программный	1.08
sang	0.70	сопрано	0.97	oracle	0.82	пользователь	1.04
singing	0.69	вагнер	0.90	enterprise	0.78	обеспечение	1.02
operas	0.68	оркестр	0.82	users	0.78	система	0.96

Рис. 13: Примеры тем из двуязычной тематической модели Википедии. Показаны первые 10 слов каждой темы и их вероятности $p(w|t)$ в процентах.

переводом слова w из языка ℓ, то их распределения тем $p(t|u)$ и $p(t|w)$ должны быть близки в смысле KL-дивергенции:

$$R(\Phi) = \sum_{w,u} \sum_{t \in T} n_{ut} \ln \varphi_{wt}.$$

Согласно формуле M-шага, вероятность слова в теме увеличивается, если оно имеет переводы, имеющие высокую вероятность в данной теме:

$$\varphi_{wt} = \underset{w \in W^\ell}{\text{norm}}\Big(n_{wt} + \tau \sum_u n_{ut}\Big).$$

Этот регуляризатор не учитывал, что перевод слова может зависеть от темы, и что среди переводов слова могут находиться переводы его омонимов. Поэтому в той же работе был предложен второй регуляризатор, который вводил в модель новые параметры $\pi_{uwt} = p(u|w,t)$ — вероятности того, что слово u является переводом слова w в теме t. Предполагается, что тема t, как распределение $\hat{p}(u|t) = \frac{n_{ut}}{n_t}$ над словами языка k, должна быть близка в смысле KL-дивергенции к вероятностной модели той же темы $p(u|t) = \sum_w \pi_{uwt}\varphi_{wt}$, построенной по переводам слов из языка ℓ:

$$R(\Phi, \Pi) = \tau \sum_{u \in W^k} \sum_{t \in T} n_{ut} \ln \sum_{w \in W^\ell} \pi_{uwt}\varphi_{wt} \;\to\; \max_{\Phi,\Pi}.$$

Формула M-шага теперь учитывает вероятности переводов π_{uwt}. Кроме того, добавляется рекуррентная формула для оценивания этих вероятностей:

$$\varphi_{wt} = \underset{w \in W^\ell}{\text{norm}}\Big(n_{wt} + \tau \sum_u \pi_{wut} n_{ut}\Big);$$

$$\pi_{uwt} = \underset{u \in W^k}{\text{norm}}\big(\pi_{wut} n_{ut}\big).$$

Темы, в которых $p(\text{«sum»}|\text{«сумма»}, t) > 0.9$

Тема №6		Тема №12		Тема №20	
множество	set	математика	triangle	вектор	vector
пространство	space	треугольник	square	координата	coordinate
группа	point	теорема	number	пространство	field
точка	left	точка	point	преобразование	tensor
элемент	limit	математический	theorem	базис	transform
функция	symmetry	угол	angle	тензор	basis
предел	function	координата	mathematics	сила	space
отображение	open	экономика	real	векторный	force
симметрия	property	число	theory	точка	rotation
открытый	topology	квадрат	geometry	система	thermometer

Темы, в которых $p(\text{«total»}|\text{«сумма»}, t) > 0.9$

Тема №5		Тема №19		Тема №22	
орбита	space	программный	software	игра	game
аппарат	nasum	версия	version	видеосигнал	character
космический	orbit	работа	news	игрок	video
земля	instrument	компания	company	фильм	player
поверхность	earth	анонимный	work	головоломка	series
солнечный	surface	примечание	note	серия	puzzle
станция	solar	терминатор	release	качество	movie
запуск	system	журнал	support	шахматы	jason
система	landing	рей	terminator	джейсон	world
атмосфера	camera	персонаж	anonymous	буква	chess

Рис. 14: Примеры тем, в которых слово «сумма» имеет разные переводы.

Связывание параллельных текстов сильнее улучшает качество поиска, чем оба способа учёта словарей [10]. Второй способ немного лучше первого. Кроме того, он позволяет выбирать варианты перевода в зависимости от контекста, что может быть полезно для статистического машинного перевода, рис. 14.

Модальности категорий и авторов. Допустим, что распределения тем в документах $p(t\,|\,d)$ порождаются одной из модальностей, например, авторами, рубриками или категориями. Будем считать, что с каждым термом w в каждом документе d связана не только тема $t \in T$, но и категория c из заданного множества категорий C. Расширим вероятностное пространство до множества $D \times W \times T \times C$. Пусть известно подмножество категорий $C_d \subseteq C$, к которым может относиться документ d.

Рассмотрим мультимодальную тематическую модель (38), в которой распределение вероятности тем документов $\theta_{td} = p(t\,|\,d)$ описывается смесью распределений тем категорий $\psi_{tc} = p(t\,|\,c)$ и категорий документов $\pi_{cd} = p(c\,|\,d)$:

$$p(w\,|\,d) = \sum_{t\in T} p(w\,|\,t) \sum_{c\in C_d} p(t\,|\,c)p(c\,|\,d) = \sum_{t\in T}\sum_{c\in C_d} \varphi_{wt}\psi_{tc}\pi_{cd}. \tag{44}$$

Это также задача стохастического матричного разложения, только теперь требуется найти три матрицы: Φ — матрица термов тем, $\Psi = (\psi_{tc})_{T\times C}$ — *матрица тем категорий*, $\Pi = (\pi_{cd})_{C\times D}$ — *матрица категорий документов*.

Модель основана на двух гипотезах условной независимости:

$p(t\,|\,c,d) = p(t\,|\,c)$ — тематика документа d зависит не от самого документа, а только от того, каким категориям он принадлежит;

$p(w\,|\,t,c,d) = p(w\,|\,t)$ — распределение термов определяется только тематикой документа и не зависит от его термов и категорий.

Кроме того, предполагается, что $\pi_{cd} = p(c\,|\,d) = 0$ для всех $c \notin C_d$.

Задача максимизации регуляризованного правдоподобия:

$$\sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{t\in T}\sum_{c\in C_d} \varphi_{wt}\psi_{tc}\pi_{cd} + R(\Phi,\Psi,\Pi) \;\to\; \max_{\Phi,\Psi,\Pi}; \tag{45}$$

$$\sum_{w\in W} \varphi_{wt} = 1,\ \varphi_{wt} \geqslant 0; \quad \sum_{t\in T} \psi_{tc} = 1,\ \psi_{tc} \geqslant 0; \quad \sum_{c\in C_d} \pi_{cd} = 1,\ \pi_{cd} \geqslant 0. \tag{46}$$

Теорема 9.2. *Пусть функция $R(\Phi,\Psi,\Pi)$ непрерывно дифференцируема. Точка локального экстремума (Φ,Ψ,Π) задачи (45), (46) удовлетворяет системе уравнений со вспомогательными переменными $p_{tcdw} = p(t,c\,|\,d,w)$, если из решения исключить нулевые столбцы матриц Φ, Ψ, Π:*

$$p_{tcdw} = \underset{(t,c)\in T\times C_d}{\operatorname{norm}} \big(\varphi_{wt}\psi_{tc}\pi_{cd}\big);$$

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

$$\psi_{tc} = \underset{t\in T}{\operatorname{norm}}\left(n_{tc} + \psi_{tc}\frac{\partial R}{\partial \psi_{tc}}\right); \qquad n_{tc} = \sum_{d\in D}\sum_{w\in d} n_{dw}p_{tcdw};$$

$$\pi_{cd} = \underset{c\in C_d}{\operatorname{norm}}\left(n_{cd} + \pi_{cd}\frac{\partial R}{\partial \pi_{cd}}\right); \qquad n_{cd} = \sum_{w\in d}\sum_{t\in T} n_{dw}p_{tcdw}.$$

Доказательство опирается на лемму 3.2 о максимизации на единичных симплексах и проводится аналогично доказательству теоремы 4.1.

Модель трёхматричного разложения наиболее известна как *автор-тематическая модель* ATM (author-topic model), в которой порождающей модальностью являются авторы документов [155]. В *тематической модели тегирования документов* TWTM (tag weighted topic model) порождающей модальностью являются теги документа [103]. Аналогичная модель использовалась для обработки видеопотоков в [78]: документы d соответствовали последовательным 1-секундным видеоклипам, термы w — элементарным визуальным событиям, темы t — действиям, состоящим из сочетания событий, категории c — более сложным поведениям, состоящим из сочетания действий, причём ставилась задача выделить в каждом клипе одно основное поведение.

Модель (44) можно упростить и свести снова к двухматричному разложению, если отождествить темы с категориями, $C \equiv T$, и взять единичную матрицу Ψ. Данная модель известна в литературе как Flat-LDA [156] и Labeled-LDA [151]. Её выразительные возможности беднее, чем у PLSA и LDA, так как значительная доля элементов матрицы $\Pi \equiv \Theta$ фиксированы и равны нулю.

Трёхматричные разложения пока не реализованы в библиотеке `BigARTM`.

Модальность времени и темпоральные модели. Время создания документов важно при анализе новостных потоков, научных публикаций, патентных баз, данных социальных сетей. Тематические модели, учитывающие время, называются *темпоральными*. Они позволяют выделять событийные и перманентные темы, детектировать новые темы, прослеживать развитие тем во времени, выделять тренды.

Пусть I — конечное множество интервалов времени, и каждый документ относится к одному или нескольким интервалам, D_i — подмножество документов, относящихся к интервалу i. Будем полагать, что темы как распределения $p(w\,|\,t)$ не меняются во времени. Требуется найти распределение каждой темы во времени $p(i\,|\,t)$.

Тривиальный подход заключается в том, чтобы построить тематическую модель без учёта времени, затем найти распределение тем в каждом интервале $p(t\,|\,i)$ как среднее θ_{td} по всем документам $d \in D_i$ и перенормировать условные вероятности: $p(i\,|\,t) = p(t\,|\,i)\frac{p(i)}{p(t)}$. Недостаток данного подхода в том, что информация о времени никак не используется при обучении модели и не влияет на формирование тем.

В ARTM эта проблема решается введением модальности времени I. Искомое распределение $p(i\,|\,t) = \varphi_{it}$ получается в столбце матрицы Φ. Дополнительные ограничения на поведение тем во времени можно вводить с помощью регуляризации.

В одной из первых темпоральных тематических моделей TOT (topics over time) [196] каждая тема моделировалась параметрическим β-распределением во времени. Это семейство монотонных и унимодальных непрерывных функций, с помощью которого можно описывать узкие пики событийных тем и ограниченный набор трендов. Темы с несколькими всплесками данная модель описывает плохо.

Непараметрические темпоральные модели способны описывать произвольные изменения тем во времени. Рассмотрим два естественных предположения и формализуем их с помощью регуляризации.

Во-первых, предположим, что многие темы являются событийными и имеют относительно небольшое «время жизни», поэтому в каждом интервале времени i присутствуют не все темы. Потребуем разреженности распределений $p(t\,|\,i)$ с помощью кросс-энтропийного регуляризатора:

$$R_{\text{разр}}(\Phi \text{ или } \Theta) = -\tau_{\text{разр}} \sum_{i\in I} \sum_{t\in T} \ln p(t\,|\,i).$$

Во-вторых, предположим, что распределения $p(i\,|\,t)$ как функции времени меняются не слишком быстро. Для этого введём регуляризатор сглаживания, минимизирующий модули разностей $p(i\,|\,t)$ в соседних интервалах времени:

$$R_{\text{сгл}}(\Phi \text{ или } \Theta) = -\tau_{\text{сгл}} \sum_{i\in I} \sum_{t\in T} \big|p(i\,|\,t) - p(i{-}1\,|\,t)\big|.$$

Такой регуляризатор принято использовать при интерполяции разрывных временных рядов [91]. Модуль разности (L_1-норма), в отличие от квадрата разности (L_2-нормы), не штрафует модель временного ряда за резкие скачки. Эта особенность полезна при моделировании тематики новостных потоков. Произошедшее событие вызывает скачкообразный всплеск вероятности $p(i\,|\,t)$ для новой темы t, которая до этого оставалась равной нулю. L_1-регуляризатор не пытается сглаживать такие скачки, в отличие от L_2-регуляризатора. Использование L_2-нормы может приводить к нелепым результатам при обработке исторических данных — «челябинский метеорит» ещё не прилетел, а некоторая доля его темы уже появилась в потоке новостей.

Оба регуляризатора можно записать и как функцию от Φ, и как функцию от Θ. В первом случае вводится модальность интервалов времени. Во втором случае при обработке документов, относящихся к интервалу i, придётся обеспечивать доступ

к вектор-столбцам θ_d документов, относящихся к соседним интервалам $i \pm 1$. Технически это менее удобно, поэтому будем рассматривать только первый вариант.

Рассмотрим в более общем виде задачу мультимодального тематического моделирования с взвешенной суммой негладких L_1-регуляризаторов $|R_j(\Phi,\Theta)|$, $j \in J$:

$$\sum_{m\in M} \tau_m \sum_{d\in D} \sum_{w\in W^m} n_{dw} \ln \sum_{t\in T} \varphi_{wt}\theta_{td} + R(\Phi,\Theta) - \sum_{j\in J} \big|R_j(\Phi,\Theta)\big| \;\to\; \max_{\Phi,\Theta}, \tag{47}$$

при обычных ограничениях неотрицательности и нормировки (40).

Теорема 9.3. *Пусть функции $R(\Phi,\Theta)$ и $R_j(\Phi,\Theta)$ непрерывно дифференцируемы. Тогда точка (Φ,Θ) локального экстремума задачи (47) с ограничениями (40) удовлетворяет системе уравнений со вспомогательными переменными $p_{tdw} = p(t\,|\,d,w)$, если из решения исключить нулевые столбцы матриц Φ, Θ:*

$$p_{tdw} = \underset{t\in T}{\operatorname{norm}}\big(\varphi_{wt}\theta_{td}\big); \tag{48}$$

$$\varphi_{wt} = \underset{w\in W^m}{\operatorname{norm}}\left(n_{wt} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}} - \varphi_{wt}\sum_{j\in J}\operatorname{sign}(R_j)\frac{\partial R_j}{\partial \varphi_{wt}}\right); \tag{49}$$

$$\theta_{td} = \underset{t\in T}{\operatorname{norm}}\left(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}} - \theta_{td}\sum_{j\in J}\operatorname{sign}(R_j)\frac{\partial R_j}{\partial \theta_{td}}\right); \tag{50}$$

где $\operatorname{sign}(x) = [x>0] - [x<0]$, *выражения для* n_{wt} *и* n_{td} *те же, что в* (42)–(43).

В силу негладкости функционала лемма о максимизации на единичных симплексах непосредственно неприменима. Эта проблема обходится введением дополнительных неотрицательных переменных ρ_j^+ и ρ_j^-, удовлетворяющих системе ограничений $|R_j(\Phi,\Theta)| = \rho_j^+ + \rho_j^-$ и $R_j(\Phi,\Theta) = \rho_j^+ - \rho_j^-$. В результате регуляризованный критерий становится гладким, что позволяет применить к задаче условия Каруша–Куна–Таккера. Доказательство можно найти в работе Н. В. Дойкова[2].

При введении регуляризатора $R_{\text{сгл}}(\Phi)$ формула М-шага для модальности времени принимает следующий вид:

$$\varphi_{it} = \underset{i\in I}{\operatorname{norm}}\Big(n_{it} - \tau_{\text{сгл}}\varphi_{it}\big(\operatorname{sign}(\varphi_{it} - \varphi_{i-1,t}) + \operatorname{sign}(\varphi_{it} - \varphi_{i+1,t})\big)\Big). \tag{51}$$

Регуляризатор $R_{\text{сгл}}$ сглаживает временной ряд $\varphi_{it} = p(i\,|\,t)$ в каждой точке по отношению к соседним точкам слева и справа. Если значение φ_{it} становится выше обоих соседних $\varphi_{i\pm1,t}$, то регуляризатор уменьшает его; если ниже, то, наоборот, увеличивает. Если значение φ_{it} попадает между ними, то оно не изменяется.

[2] Дойков Н. В. Адаптивная регуляризация вероятностных тематических моделей. Выпускная квалификационная работа бакалавра, ВМК МГУ, 2015.
`http://www.MachineLearning.ru/wiki/images/9/9f/2015_417_DoykovNV.pdf`

Выводы по главе

- Введение модальностей — простое, но мощное обобщение тематических моделей. Оно позволяет использовать текстовые данные совместно с разнотипной дополнительной информацией.
- Мультиязычные и темпоральные модели являются частными случаями мультимодальных, но могут требовать дополнительной специфичной регуляризации.
- В трёхматричных тематических моделях с авторами или категориями не только вводятся новые модальности, но и меняется структура матричного разложения.
- Во всех этих модификациях EM-алгоритм легко выводится по лемме о максимизации на единичных симплексах.
- Веса модальностей, как и коэффициенты регуляризации, приходится подбирать в экспериментах по выбранным критериям.
- Далее мы увидим ещё много примеров использования модальностей для формализации требований к тематической модели.

10 Моделирование транзакционных данных

Обычные тематические модели текстовых коллекций описывают вхождения слов в документы. Мультимодальные модели описывают документы, в которых содержатся термы различных модальностей: слова, теги, 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 (отношения, взаимосвязи, взаимодействия) 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 дата и время транзакции.

Во всех приведённых примерах транзакция нескольких объектов не может быть сведена к парным взаимодействиям. Попытка исключения любого объекта из транзакции приводит к утрате важной части информации. Однако возможны и такие ситуации, когда транзакция распадается на пары. Например, в системе рекомендаций музыки транзакция «трек r исполнителя a находится в альбоме d, вышедшем в году y» описывается, казалось бы, четвёркой объектов (r, a, d, y). Однако она распадается на парные взаимосвязи (d, r), (d, a), (d, y), в которых альбомы играют роль документов. Такие данные могут быть описаны обычной мультимодальной моделью.

Для тематического моделирования транзакционных данных удобно понятие гиперграфа. Гиперграф обобщает понятие графа и отличается от него тем, что рёбрами в нём могут быть не только пары вершин, но и подмножества из трёх и более вершин. Вершины гиперграфа соответствуют термам различных модальностей, рёбра — транзакциям. Задача заключается в том, чтобы по наблюдаемой выборке транзакций восстановить неизвестные тематические распределения вершин $p(t|v)$. Предполагается, что вероятность транзакции тем выше, чем более схожи тематики её вершин.

В проекте `BigARTM` реализована описанная ниже гиперграфовая тематическая модель транзакционных данных.

Тематические модели на гиперграфах. *Гиперграф* $\Gamma = \langle V, E \rangle$ определяется множеством вершин-термов V и множеством рёбер (транзакций) E. Каждое ребро e из E образуется подмножеством вершин, $e \subset V$.

Каждая вершина $v \in V$ имеет *модальность* $m = \mu(v)$ из конечного множества модальностей M. Множество всех вершин разбивается на непересекающиеся подмножества по модальностям:

$$V = \bigsqcup_{m \in M} V_m, \qquad V_m = \{v \in V \colon \mu(v) = m\}.$$

Например, 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

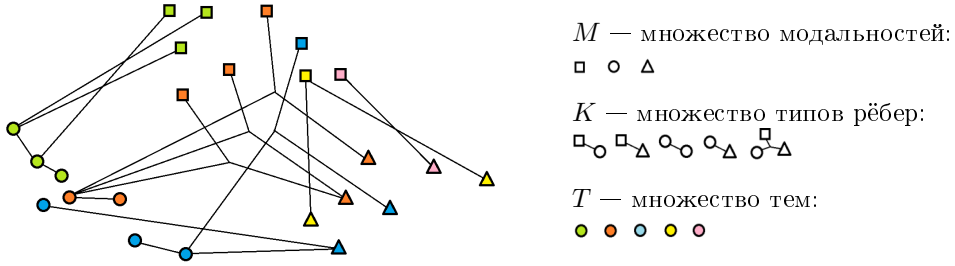


Рис. 15: Пример гиперграфа с вершинами трёх модальностей, рёбрами-транзакциями пяти типов и пятью темами.

по объявлениям b на страницах s, могут иметься данные о посещениях страниц пользователями (u, s), 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 E_k независимых наблюдений $(e, t) \in 2^V \times T$, порождаемая распределением $p_k(e, t)$, своим для каждого типа $k \in K$. 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. 15 показан пример гиперграфа с вершинами трёх модальностей, рёбрами-транзакциями пяти типов и пятью темами.

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 e, за исключением вершины-контейнера d. Аналогично документу, с контейнером связано распределение тем $p(t\,|\,d)$. Множество всех вершин-контейнеров обозначим через D.

Далее 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. Казалось бы, на практике это предположение может не выполняться. Например, распределения слов в текстах веб-страниц, в пользовательских запросах и в рекламных баннерах могут значимо различаться для одной и той же темы. Однако это ограничение нетрудно обойти, если построить модель с тремя разными модальностями слов для этих трёх типов транзакций. Сделать эти распределения похожими можно с помощью регуляризации.

Наконец, введём гипотезу условной независимости вершин v, порождаемых одним и тем же ребром (d, x):

$$p(x\,|\,t) = \prod_{v \in x} p(v\,|\,t).$$

При сделанных допущениях 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 других вершин из своего распределения $p(v\,|\,t)$ над множеством V_m.

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

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

Параметрами этой модели являются условные вероятности вершин в темах $\varphi_{vt} = p(v\,|\,t)$, нормированные по каждой модальности $v \in V_m$, и условные вероятности тем в контейнерах $\theta_{td} = p(t\,|\,d)$. В матричных обозначениях параметрами являются матрицы Φ_m, $m \in M$ и Θ, как и в случае мультимодальной тематической модели (39).

Гиперграфовая модель является широким обобщением обычных тематических моделей. В частности, она переходит в PLSA в случае, когда модальностей две — документы $V_1 = D$ и термы $V_2 = W$, тип рёбер только один — пары $(d, w) \in D \times W$, в которых документы d являются контейнерами.

Гиперграфовый EM-алгоритм. Для оценивания параметров модели применим принцип максимума правдоподобия. Будем максимизировать сумму логарифмов правдоподобия по всем типам рёбер k с весами τ_k и регуляризатором $R(\Phi, \Theta)$:

$$\sum_{k\in K} \tau_k \sum_{dx\in E_k} n_{kdx} \ln\Bigg(\sum_{t\in T} \theta_{td} \prod_{v\in x} \varphi_{vt}\Bigg) + R(\Phi,\Theta) \to \max_{\Phi,\Theta}; \tag{52}$$

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

Теорема 10.1. *Пусть функция $R(\Phi, \Theta)$ непрерывно дифференцируема. Точка локального максимума (Φ, Θ) задачи (52) 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:*

$$p_{tdx} = \underset{t\in T}{\operatorname{norm}}\Bigg(\theta_{td} \prod_{v\in x} \varphi_{vt}\Bigg); \tag{53}$$

$$\varphi_{vt} = \underset{v\in V_m}{\operatorname{norm}}\Bigg(n_{vt} + \varphi_{vt}\frac{\partial R}{\partial \varphi_{vt}}\Bigg); \qquad n_{vt} = \sum_{k\in K}\sum_{dx\in E_k} [v \in x]\,\tau_k n_{kdx} p_{tdx}; \tag{54}$$

$$\theta_{td} = \underset{t\in T}{\operatorname{norm}}\Bigg(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\Bigg); \qquad n_{td} = \sum_{k\in K}\sum_{dx\in E_k} \tau_k n_{kdx} p_{tdx}. \tag{55}$$

Доказательство. Воспользуемся леммой 3.2 о максимизации на единичных симплексах, выделив вспомогательные переменные p_{tdx}, определённые в (53):

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

Теорема доказана.

Типы транзакций и их весовые коэффициенты. В описанной модели нет никаких ограничений на то, каким образом транзакции группируются по типам. Любая транзакция, независимо от типа, может содержать любое число термов любых модальностей. На практике удобно относить к одному типу все транзакции, имеющие общее происхождение и структуру. Однако формально такого ограничения нет.

Весовые коэффициенты τ_k позволяют сбалансировать модель с учётом количества транзакций в выборках $|E_k|$. Чем больше τ_k, тем сильнее транзакции типа k повлияют на модель.

В некоторых задачах определённые типы транзакций не должны влиять на тематику контейнеров. Например, в мультиязычной модели может быть выбран один главный язык, определяющий тематику многоязычного документа, однако темы должны получить свои распределения слов в каждом языке. В таком случае для слов главного языка вводится весовой коэффициент τ_k, на порядки превышающий весовые коэффициенты слов остальных языков. Другой пример: в рекомендательной системе музыкальные треки должны влиять на темы, если они находятся в плей-листах пользователей, но не в альбомах исполнителей. Тематика альбомов должна определяться исходя из предпочтений пользователей, но не наоборот. Поэтому веса треков в плей-листах должны на порядки превышать веса треков в альбомах.

Гиперграфовые модели для рекомендательных систем. Пусть имеется конечное множество клиентов или пользователей U (users) и конечное множество объектов или товаров I (items), относительно которых у клиентов могут быть различные предпочтения, вкусы или интересы. Вероятностная тематическая модель рекомендательной системы предсказывает предпочтения клиентов:

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

Она эквивалентна тематической модели текстовой коллекции с точностью до замены терминологии: «документы $\to$ клиенты», «слова $\to$ объекты», «темы $\to$ интересы». Исходными данными являются счётчики n_{ui}, описывающие частоту использования объекта i клиентом u. В зависимости от приложения это могут быть покупки, посещения, обращения, лайки и т. д.

В рекомендательных системах существует проблема «холодного старта»: новому клиенту нечего порекомендовать, поскольку мы не имеем 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 бинарные данные n_{ua} о наличии у клиента u атрибута a из заданного конечного множества A или данные n_{ib} о наличии у объекта i свойства b из заданного конечного множества B. Например, объекты могут иметь текстовые описания, и тогда B — это словарь терминов, используемых в этих описаниях. Такие рекомендательные системы называются, соответственно, учитывающими атрибуты (attribute-aware) и учитывающими контент (content-aware). Они позволяют определять тематические векторы клиентов и объектов даже тогда, когда по ним нет основных данных о предпочтениях n_{ui}. В качестве дополнительных данных могут также использоваться советы клиентов друг другу. Это попарные взаимодействия между клиентами $n_{uu'}$ или данные о доверии (trust-aware). Итак, в задаче

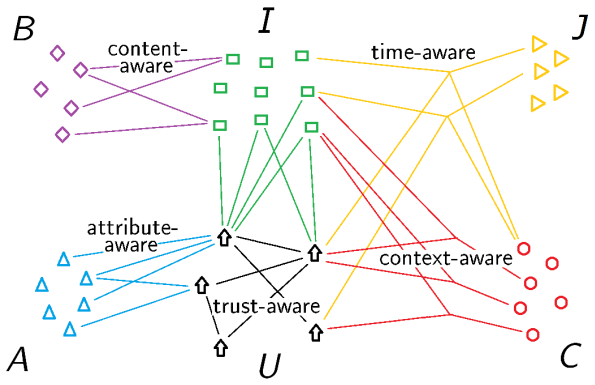


Рис. 16: Транзакционные данные в рекомендательной системе описывают взаимодействия между шестью модальностями: клиенты U, объекты I, атрибуты клиентов A, свойства объектов B, ситуативные контексты C, интервалы времени J.

появляется уже четыре модальности U, I, A, B, между элементами которых наблюдаются парные взаимодействия. Возможны и тройные взаимодействия, например, n_{uib} — клиент u тегировал объект i свойством b.

Предпочтения клиентов могут изменяться со временем или зависеть от ситуативного контекста: покупка для себя, для офиса или в подарок; просмотр фильма с друзьями, с подругой или с детьми и т. д. Для учёта такой информации вводятся ещё две модальности: множество ситуаций C и множество интервалов времени J. Взаимодействия между ними могут описываться транзакциями из трёх или четырёх 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), см. рис. 16.

Симметризованные гиперграфовые модели подходят для задач, в которых содержимое контейнеров может изменяться. В текстовых коллекциях документы появляются целиком и в дальнейшем не меняются. Совершенно иная ситуация в рекомендательных системах. Транзакция «пользователь u выбрал товар i» на первый взгляд аналогична транзакции «документ d содержит слово w». Различие в том, что документ статичен, тогда как у каждого пользователя может расти множество выбранных товаров, а у каждого товара — множество выбравших его пользователей. Кто в таком случае должен играть роль контейнера — товар или пользователь? То и другое неудобно с точки зрения пакетного EM-алгоритма; гораздо естественнее было бы формировать пакеты по времени поступления данных.

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:

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

Возможен также вариант, когда тема t порождается из распределения $\pi_{kt} = p_k(t)$, общего для всех транзакций данного типа k. Мы не будем рассматривать этот случай, поскольку он легко расписывается по аналогии.

Пусть E_k — наблюдаемая выборка рёбер-транзакций типа k, n_{kx} — число наблюдений ребра x в выборке E_k. Выпишем задачу максимизации регуляризованного правдоподобия при обычных ограничениях нормировки и неотрицательности:

$$\sum_{k\in K}\tau_k\sum_{x\in E_k}n_{kx}\ln\Big(\sum_{t\in T}\pi_t\prod_{v\in x}\varphi_{vt}\Big)\;+\;R(\Phi,\pi)\to\max_{\Phi,\pi};\tag{56}$$

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

Теорема 10.2. *Пусть функция $R(\Phi,\pi)$ непрерывно дифференцируема. Точка локального максимума (Φ,π) задачи (56) 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:*

$$p_{tx}=\underset{t\in T}{\operatorname{norm}}\Big(\pi_t\prod_{v\in x}\varphi_{vt}\Big).\tag{57}$$

$$\varphi_{vt}=\underset{v\in V_m}{\operatorname{norm}}\Big(n_{vt}+\varphi_{vt}\frac{\partial R}{\partial\varphi_{vt}}\Big);\qquad n_{vt}=\sum_{k\in K}\sum_{x\in E_k}[v\in x]\,\tau_k n_{kx}p_{tx};\tag{58}$$

$$\pi_t=\underset{t\in T}{\operatorname{norm}}\Big(n_t+\pi_t\frac{\partial R}{\partial\pi_t}\Big);\qquad n_t=\sum_{k\in K}\sum_{x\in E_k}\tau_k n_{kx}p_{tx}.\tag{59}$$

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

В BigARTM такая модель непосредственно не реализована, но её нетрудно с хорошей точностью симулировать. Для этого симметризованные транзакции, не имеющие контейнера, собираются в псевдо-контейнеры d произвольным образом; их счётчики n_{td} аккумулируются: $n_t=\sum_d n_{td}$, и накопленные суммы используются для сильного сглаживания всех столбцов матрицы Θ в сторону общего вектора (n_t).

Транзакции с главными и подчинёнными термами. Предположим, что каждая транзакция состоит из двух непересекающихся подмножеств термов: $x\sqcup x'$. Подмножество *главных термов* x определяет тематику транзакции $p(t\,|\,x,x')=p(t\,|\,x)$, которая затем передаётся подмножеству *подчинённых термов* x'. Это предположение можно вводить как для симметризованных моделей с транзакциями вида $e=(x,x')$, так и для моделей с контейнерными транзакциями $e=(d,x,x')$.

Например, к банковской транзакции «покупатель b купил у продавца s товар g» может прилагаться текст платёжного поручения. Он не должен влиять на тематику транзакции, которая определяется экономической деятельностью продавца и покупателя. Термы b,s,g являются главными, а слова из платёжного поручения — подчинёнными. Просто игнорировать эти тексты не хотелось бы, так как распределения слов $p(w\,|\,t)$ полезны для содержательной интерпретации тем как видов деятельности компаний. Заметим также, что банковские транзакции не являются контейнерными, поскольку контрагенты постоянно осуществляют новые сделки.

Сделаем ещё один шаг в обобщении гипеграфовой тематической модели. Будем полагать, что коллекция E содержит как контейнерные, так и симметризованные транзакции. Наличие контейнера d у транзакции будем обозначать условием $[e=(d,x,x')]$, а отсутствие контейнера — условием $[e=(x,x')]$.

Модификация EM-алгоритма объединяет контейнерный вариант (53)–(55) и симметризованный вариант (57)–(59):

$$p_{te} = \underset{t\in T}{\operatorname{norm}}\left(\Big([e=(d,x,x')]\theta_{td} + [e=(x,x')]\pi_t\Big)\prod_{v\in x}\varphi_{vt}\right);$$

$$\varphi_{vt} = \underset{v\in V_m}{\operatorname{norm}}\left(n_{vt} + \varphi_{vt}\frac{\partial R}{\partial \varphi_{vt}}\right); \qquad n_{vt} = \sum_{k\in K}\sum_{e\in E_k}[v\in e]\,\tau_k n_{ke} p_{te};$$

$$\theta_{td} = \underset{t\in T}{\operatorname{norm}}\left(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right); \qquad n_{td} = \sum_{k\in K}\sum_{e\in E_k}[e=(d,x,x')]\tau_k n_{ke} p_{te};$$

$$\pi_t = \underset{t\in T}{\operatorname{norm}}\left(n_t + \pi_t\frac{\partial R}{\partial \pi_t}\right); \qquad n_t = \sum_{k\in K}\sum_{e\in E_k}[e=(x,x')]\tau_k n_{ke} p_{te}.$$

Основная модификация касается E-шага: если транзакция e относится к контейнеру d, то используется распределение $p(t\,|\,d) = \theta_{td}$, если же она не относится ни к одному контейнеру, то используется распределение $p(t) = \pi_t$, общее для всех транзакций. В обоих вариантах в результате получается распределение тем для транзакции $p(t\,|\,e) = p_{te}$, которое затем используется во всех формулах M-шага.

Гиперграфовые языковые модели. В текстах естественного языка ребром гиперграфа (транзакцией) можно считать любое подмножество слов, предположительно генерируемых одной общей темой из распределения $p(t\,|\,d)$. Это могут быть слова, связанные синтаксически — предложение, фраза, словосочетание, синтагма. Либо слова, связанные лексически, например, тезаурусными отношениями [14] синонимии, «часть–целое», «общее–частное». Использование такого рода связей можно считать уходом от гипотезы «мешка слов» — наименее реалистичного и наиболее критикуемого предположения в вероятностном тематическом моделировании.

Мы вернёмся к этой идее в главе 15, когда будем рассматривать тематические модели связного текста. Здесь же отметим естественную языковую интерпретацию транзакций с главными и подчинёнными термами. Главные термы определяют тематику транзакции (например, предложения). В роли подчинённых могут выступать слова общей лексики или редкие слова. Более того, способ разбиения каждой транзакции $e = (d, x, x')$ на подмножества x и x' в транзакционной модели не фиксирован и может уточняться в ходе итераций. Это открывает массу интересных возможностей для моделирования тематики слов внутри предложений.

Выводы по главе

- Гиперграфовые тематические модели — это мощное обобщение мультимодальных моделей. Ещё сильнее расширяется класс решаемых прикладных задач.
- Новый общий взгляд на тематическое моделирование: это векторизация вершин графа или гиперграфа по наблюдаемой выборке его рёбер. Обычный текст — это двудольный граф, в котором вершины-слова соединены рёбрами с вершинами-документами.
- Лемма о максимизации на единичных симплексах даже в таком, казалось бы довольно сложном, обобщении позволила легко вывести EM-алгоритм.

11 Моделирование зависимостей

Тематическая модель формирует векторное описание документа $p(t|d)$, которое может быть использовано для предсказательного моделирования, в частности, для решения задач классификации и регрессии на текстах. Классификация (или категоризация) реализуется особенно просто, если классы считать модальностью.

Регрессия на текстах показывает пример интересного приёма, когда дополнительные параметры модели (коэффициенты линейной модели регрессии) пересчитываются итерационно после каждого прохода коллекции. Очень похожий приём используется и в модели CTM, которая выявляет парные корреляции между темами.

Техника числовых модальностей используется в том случае, когда с текстом связана не одна числовая величина, которую необходимо предсказать, а числовая последовательность, порождаемая смесью непрерывных вероятностных распределений.

Классификация. Рассмотрим коллекцию документов с модальностями термов W и классов C. Для каждого документа d известно подмножество классов $C_d \subset C$. Требуется классифицировать новые документы с неизвестными C_d. Для этого будем использовать *линейную вероятностную модель классификации* документа по его тематическому вектору $\theta_{td} = p(t|d)$:

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

Документ d относится к классу c, если $p(c|d) \geqslant \gamma_c$.

Коэффициенты линейной модели $\varphi_{ct} = p(c|t)$ и пороги γ_c обучаются по выборке документов с известными C_d. Вектор нового документа θ_d вычисляется тематической моделью только по его термам.

Тематическая модель классификации Dependency LDA [156] является байесовским аналогом мультимодальной модели (38). Эксперименты в [156] показали, что тематические модели превосходят обычные методы многоклассовой классификации на больших текстовых коллекциях с большим числом несбалансированных, пересекающихся, взаимозависимых классов. В [178] 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 (multilabel classification) документ может относиться как к одному классу, так и к большому числу классов. *Взаимозависимые* классы имеют общие термы и темы, поэтому при классификации документа могут вступать в конкуренцию.

Регуляризация по отрицательным примерам использует данные о том, что документ d из обучающей выборки не принадлежит подмножеству классов $C'_d \subset C$. В этом случае запишем правдоподобие выборки для задачи бинарной классификации:

$$L(\Phi, \Theta) = \sum_{d \in D} \sum_{c \in C_d} \ln \sum_{t \in T} \varphi_{ct} \theta_{td} + \tau \sum_{d \in D} \sum_{c \in C'_d} \ln\Big(1 - \sum_{t \in T} \varphi_{ct} \theta_{td}\Big) \;\to\; \max.$$

Первое слагаемое есть log-правдоподобие модальности классов (38), если положить $n_{dc} = [c \in C_d]$. Второе слагаемое можно рассматривать как регуляризатор отрицательных примеров, построенный по данным о не-принадлежности документов классам. Коэффициент регуляризации τ можно полагать равным единице.

Частотная регуляризация (label regularization) хорошо зарекомендовала себя в задачах с несбалансированными классами [111, 156]. Потребуем, чтобы оценка безусловного распределения классов по коллекции $p(c) = \sum_t \varphi_{ct} p(t)$ была близка к наблюдаемым частотам классов $\hat{p}(c) = \frac{1}{|D|}|D_c|$, где $D_c = \{d \in D \colon c \in C_d\}$ — множество документов, относящихся к классу c. Выразим данное требование с помощью сглаживающего регуляризатора кросс-энтропии, который можно интерпретировать и как максимизацию правдоподобия для модели дискретного распределения классов $p(c)$:

$$R(\Phi) = \tau \sum_{c \in C} |D_c| \ln \sum_{t \in T} n_t \varphi_{ct} \to \max,$$

где $n_t = \sum_c n_{ct}$ — число термов модальности C, относящихся к теме t во всей коллекции. Подставляя этот регуляризатор в (17), получим формулы M-шага:

$$\varphi_{ct} = \underset{w \in W}{\operatorname{norm}} \left(n_{ct} + \tau |D_c| \frac{n_t \varphi_{ct}}{\sum_s n_s \varphi_{cs}} \right). \tag{60}$$

Частотная регуляризация использовалась в тематической модели Prior-LDA, которая была предложена в [156] как улучшение модели Flat-LDA.

Регрессия. Задачи предсказания числовой величины как функции от текста возникают во многих приложениях электронной коммерции: предсказание рейтинга товара, фильма или книги по тексту отзыва; предсказание числа кликов по тексту рекламного объявления; предсказание зарплаты по описанию вакансии; предсказание полезности (числа лайков) отзыва на отель, ресторан, сервис. Для восстановления числовых функций по конечной обучающей выборке пар «объект–ответ» используются регрессионные модели, однако все они принимают на входе векторные описания объектов. Тематическая модель позволяет заменить текст документа d его векторным представлением θ_d. С другой стороны, критерий оптимизации регрессионной модели можно использовать в качестве регуляризатора, чтобы найти темы, наиболее информативные с точки зрения точности предсказаний [116, 167].

Пусть для каждого документа d обучающей выборки D задано целевое значение $y_d \in \mathbb{R}$. Рассмотрим *линейную модель регрессии*, которая предсказывает математическое ожидание целевой величины:

$$\mathsf{E}(y \,|\, d) = \sum_{t \in T} v_t \theta_{td},$$

где $v \in \mathbb{R}^T$ — вектор коэффициентов. Применим метод наименьших квадратов для обучения вектора v по выборке документов:

$$R(\Theta, v) = -\tau \sum_{d \in D} \Big(y_d - \sum_{t \in T} v_t \theta_{td} \Big)^2.$$

Подставляя этот регуляризатор в (18) и приравнивая нулю его производную по v, получим формулы M-шага:

$$\theta_{td} = \underset{t}{\operatorname{norm}} \Big(n_{td} + \tau v_t \theta_{td} \Big(y_d - \sum_{s \in T} v_s \theta_{sd} \Big) \Big);$$

$$v = (\Theta \Theta^T)^{-1} \Theta y.$$

Формула для вектора v является стандартным решением задачи наименьших квадратов при фиксированной матрице Θ. Вектор v можно обновлять по окончании каждого прохода коллекции, либо после обработки каждого пакета документов в онлайновом EM-алгоритме.

В [167] показано, что качество регрессии может зависеть от инициализации тематической модели, и предложено несколько методов инициализации.

На практике обычно используется более простой подход: сначала строятся тематические признаковые описания документов с помощью модели LDA, затем к этим признакам могут добавляться ещё какие-то числовые признаки текстов, и, наконец, общее признаковое описание используется для решения регрессионной задачи. Недостаток этого подхода в том, что модель LDA ничего не знает о регрессии. Регуляризация ARTM позволяет поочередно улучшать то тематическую модель с учётом регрессии, то регрессионную модель с учётом тематических признаков. В результате две модели приспосабливаются друг к другу. Векторы тем поворачиваются в пространстве таким образом, чтобы быть максимально полезными в качестве признаков регрессионной модели. Добавление дополнительных нетематических признаков в регрессионную модель в этом случае также не составляет труда.

Корреляции тем. *Модель коррелированных тем* CTM (correlated topic model) предназначена для выявления связей между темами [35]. Например, статья по геологии более вероятно связана с археологией, чем с генетикой. Знание о том, какие темы чаще совместно встречаются в документах коллекции, позволяет точнее моделировать тематику отдельных документов в мультидисциплинарных коллекциях.

Для описания корреляций удобно использовать многомерное нормальное распределение. Оно не подходит для описания неотрицательных нормированных векторстолбцов θ_d, но неплохо описывает векторы их логарифмов $\eta_{td} = \ln \theta_{td}$. Поэтому в модель вводится многомерное лог-нормальное распределение (logistic normal) с двумя параметрами: вектором математического ожидания μ и ковариационной матрицей Σ:

$$p(\eta_d \,|\, \mu, \Sigma) = (2\pi)^{-\frac{n}{2}} |\Sigma|^{-\frac{1}{2}} \exp\Big(-\tfrac{1}{2}(\eta_d - \mu)^\intercal \Sigma^{-1} (\eta_d - \mu)\Big).$$

Изначально модель CTM была разработана в рамках байесовского подхода, где возникали дополнительные технические трудности из-за того, что лог-нормальное распределение не является сопряжённым к мультиномиальному. В рамках ARTM идея CTM формализуется и реализуется намного проще.

Определим регуляризатор как логарифм правдоподобия выборки векторов документов η_d для лог-нормальной модели:

$$R(\Theta, \mu, \Sigma) = \tau \sum_{d \in D} \ln p(\eta_d \,|\, \mu, \Sigma) = -\frac{\tau}{2} \sum_{d \in D} (\ln \theta_d - \mu)^\intercal \Sigma^{-1} (\ln \theta_d - \mu).$$

Согласно (18), формула M-шага для θ_{td} принимает вид

$$\theta_{td} = \underset{t \in T}{\operatorname{norm}} \Big(n_{td} - \ \tau \sum_{s \in T} \Sigma^{-1}_{ts} \big(\ln \theta_{sd} - \mu_s \big) \Big), \tag{61}$$

где Σ^{-1}_{ts} — элементы обратной ковариационной матрицы. Параметры Σ, μ нормального распределения обновляются после каждого прохода коллекции, либо после каж-

дого пакета в онлайновом EM-алгоритме:

$$\mu = \frac{1}{|D|}\sum_{d\in D} \ln \theta_d;$$
$$\Sigma = \frac{1}{|D|}\sum_{d\in D} \big(\ln \theta_d - \mu\big)\big(\ln \theta_d - \mu\big)^{\intercal}.$$

Таким образом, трудоёмкое обращение ковариационной матрицы можно выполнять относительно редко. Чтобы получать разреженную ковариационную матрицу, в [35] использовалась LASSO-регрессия. Там же представлены примеры визуализации связей между темами, построенными по коллекции статей журнала Science за 1990–1999 гг.

Числовые модальности. В задачах регрессии с каждым текстовым документом связана одна числовая величина, которую необходимо прогнозировать. Однако бывают и такие задачи, в которых числовые величины связаны с каждым термом.

Рассмотрим тематическую модель банковских транзакционных данных. В роли документов выступают компании, термами в документе являются контрагенты — другие компании, которые заключают с данной компанией сделки по покупке или продаже товаров или услуг. Каждая сделка сопровождается текстом платёжного поручения, который содержит названия товаров или услуг. Эти названия образуют вторую модальность. Покупки и продажи рассматриваются по отдельности, что удваивает число модальностей до четырёх. Темы соответствуют видам экономической деятельности компаний. Для интерпретации тем рассматриваются два списка — товары, которые компании покупают, и товары, которые они продают, осуществляя данный вид деятельности [18].

Кроме естественных дискретных модальностей контрагентов и названий, в данной задаче имеются ещё и числовые данные об объёмах сделок, которые могут нести косвенную информацию о виде деятельности. Будем полагать, что с каждым документом d связано несколько числовых последовательностей $\{y_{dmi} \,|\, i = 1, \ldots, k_{dm}\}$, соответствующих *числовым модальностям* $m \in \tilde{M}$. В нашем случае числовых модальностей две — это объёмы сделок покупки и продажи. Предположим, что каждая тема t порождает значения y_{dmi} из распределения $p(y \,|\, t; \gamma_{tm})$ с вектором параметров γ_{tm}, которое не зависит от документа (это обычная для тематического моделирования гипотеза условной независимости). Тогда распределение значений y модальности m в документе d описывается смесью распределений:

$$p(y \,|\, d, m) = \sum_{t\in T} p(y \,|\, t; \gamma_{tm})\theta_{td}.$$

Возьмём за основу мультимодальную модель (39) с произвольным непрерывно дифференцируемым регуляризатором $R(\Phi, \Theta)$. Добавим регуляризатор числовых модальностей, определив его как взвешенную сумму логарифмов правдоподобия выборок $\{y_{dmi}\}$ с весами $\tilde{\tau}_m$:

$$\tilde{R}(\Theta, \Gamma) = \sum_{m\in\tilde{M}} \tilde{\tau}_m \sum_{d\in D} \sum_{i=1}^{k_{dm}} \ln \sum_{t\in T} p(y_{dmi} \,|\, t; \gamma_{tm})\, \theta_{td}.$$

Тогда система уравнений в теореме 9.1 преобразуется следующим образом: формулы E-шага (41) и M-шага по Φ (42) остаются в силе; к ним добавляются формулы E-шага и M-шага для числовых модальностей и изменяется формула M-шага по Θ:

$$p_{tdmi} = p(t\,|\,d, y_{dmi}) = \underset{t\in T}{\operatorname{norm}}\big(p(y_{dmi}\,|\,t;\gamma_{tm})\,\theta_{td}\big);$$

$$\gamma_{tm} = \arg\max_{\gamma\in\Gamma} \sum_{d\in D}\sum_{i=1}^{k_{dm}} p_{tdmi} \ln p(y_{dmi}\,|\,t;\gamma_{tm});$$

$$\theta_{td} = \underset{t\in T}{\operatorname{norm}}\bigg(\sum_{m\in M}\sum_{w\in W^m} \tau_m n_{dw} p_{tdw} + \sum_{m\in\tilde{M}}\sum_{i=1}^{k_{dm}} \tilde{\tau}_m p_{tdmi} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\bigg).$$

Подход на основе мультимодальной модели имеет один существенный изъян — он пренебрегает транзакционной природой данных. Каждый документ (компания) представляется «мешком названий», «мешком контрагентов», и «мешком объёмов». Однако в каждой транзакции название, контрагент и объём сделки неразрывно связаны друг с другом и порождаются общей темой (видом деятельности). Информация об этих связях игнорируется в мультимодальной модели.

Поэтому рассмотрим введение числовой модальности в гиперграфовой тематической модели, описанной в главе 10. Каждая транзакция компании d представляет собой четвёрку (d, m, x, y), где $m \in \tilde{M}$ — одна из двух числовых модальностей «объём покупки» или «объём продажи», $x \subset V$ — подмножество термов, V — множество вершин гиперграфа, полученное объединением словарей W^m всех нечисловых модальностей $m \in M$. Если компания d совершает сделку покупки, то термами в x будут компания-продавец и названия товаров, которые d у неё покупает. Если компания d совершает сделку продажи, то термами в x будут компания-покупатель и названия товаров, которые d ей продаёт. В обоих типах транзакций значение y равно объёму сделки. Тематическая модель транзакции имеет вид

$$p(x, y\,|\,d, m) = \sum_{t\in T} \theta_{td}\, p(y\,|\,t;\gamma_{tm}) \prod_{v\in x} \varphi_{vt}.$$

Задача (52) максимизации логарифма правдоподобия транзакционных данных $X = \big\{(d_i, m_i, x_i, y_i)\colon i = 1, \ldots, n\big\}$ для гиперграфовой тематической модели соответствующим образом модифицируется:

$$\sum_{(d,m,x,y)\in X} \ln\bigg(\sum_{t\in T} \theta_{td}\, p(y\,|\,t;\gamma_{tm}) \prod_{v\in x} \varphi_{vt}\bigg) + R(\Phi,\Theta) \to \max_{\Phi,\Theta,\Gamma}; \tag{62}$$

Теорема 11.1. *Пусть функция $R(\Phi,\Theta)$ непрерывно дифференцируема. Точка локального максимума (Φ,Θ,Γ) задачи (62) удовлетворяет системе уравнений относительно параметров φ_{vt}, θ_{td}, γ_{tm} и вспомогательных переменных $p_{ti} = p(t\,|\,d_i, m_i, x_i, y_i)$,*

если из решения исключить нулевые столбцы матриц Φ_m, Θ*:*

$$p_{ti} = \underset{t\in T}{\operatorname{norm}}\left(\theta_{td_i}\, p(y_i \,|\, t; \gamma_{tm_i}) \prod_{v\in x_i} \varphi_{vt}\right);$$

$$\varphi_{vt} = \underset{v\in V_m}{\operatorname{norm}}\left(n_{vt} + \varphi_{vt}\frac{\partial R}{\partial \varphi_{vt}}\right); \qquad n_{vt} = \sum_{i=1}^{n} [v \in x_i]\, p_{ti};$$

$$\theta_{td} = \underset{t\in T}{\operatorname{norm}}\left(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right); \qquad n_{td} = \sum_{i=1}^{n} [d = d_i] p_{ti};$$

$$\gamma_{tm} = \arg\max_{\gamma\in\Gamma} \sum_{i=1}^{n} [m = m_i]\, p_{ti} \ln p(y_i \,|\, t; \gamma_{tm}).$$

Интересно, что в обеих моделях, мультимодальной и гиперграфовой, нам удалось избежать конкретизации вида распределения $p(y \,|\, t; \gamma)$ до самого последнего момента. И теперь мы понимаем, каким образом в EM-алгоритм можно встроить любое предположение о виде этих распределений. Для этого достаточно уметь решать задачу максимизации взвешенного логарифма правдоподобия по $\gamma \in \Gamma$. Каждый элемент выборки, то есть каждая транзакция, имеет свой вес p_{ti}, вычисленный на E-шаге.

Это стандартная задача. Пусть для определённости $p(y \,|\, t; \gamma_{tm}) = \mathcal{N}(y; \mu_{tm}, \sigma^2_{tm})$ — многомерное нормальное распределение с математическим ожиданием μ_{tm} и дисперсией σ^2_{tm}. Тогда задача M-шага относительно параметров $\gamma_{tm} = (\mu_{tm}, \sigma^2_{tm})$ решается аналитически:

$$\mu_{tm} = \frac{\sum_{i=1}^{n} [m = m_i]\, p_{ti} y_i}{\sum_{i=1}^{n} [m = m_i]\, p_{ti}}; \qquad \sigma^2_{tm} = \frac{\sum_{i=1}^{n} [m = m_i]\, p_{ti} (y_i - \mu_{tm})^2}{\sum_{i=1}^{n} [m = m_i]\, p_{ti}}.$$

Для мультимодальной модели формулы те же, с точностью до замены p_{ti} на p_{tdmi}. В обоих случаях суммирование производится по всем транзакциям в коллекции.

Выводы по главе

- Существует много ситуаций, когда требуется моделировать зависимости между документами, термами или темами и некоторыми числовыми переменными.
- В таких задачах тематическая модель используется в качестве генератора обучаемых векторных представлений текстовых объектов, а критерий качества предсказательной модели становится регуляризатором.
- Более простой подход заключается в последовательном обучении двух моделей: сначала строится тематическая модель, затем тематические векторы используются в роли признаковых описаний объектов во второй модели. На практике это может приводить к потерям качества из-за несогласованности двух моделей.
- Преимущество ARTM в том, что при совместном обучении двух моделей тематические векторы оптимизируются под решение конечной прикладной задачи, в то же время сохраняя свойства интерпретируемости.

12 Моделирование связей между документами

Существует масса задач, в которых документы сгруппированы или связаны между собой, причём наличие взаимосвязи говорит о том, что документы имеют схожую тематику. Природа связей может быть различной: ссылки, цитирование, совместное упоминание, общие авторы или комментаторы, общие источники, близкие геолокации, и т. д.

Предположение о сходстве тематики в парах документов формализуется с помощью регуляризаторов матрицы Θ или матрицы Φ.

Ссылки и цитирование. Предположим, что если между документами c и d имеется связь, то они имеют схожие тематики $p(t\,|\,c)$ и $p(t\,|\,d)$. Формализуем это предположение с помощью регуляризатора:

$$R(\Theta) = \tau \sum_{d,c} n_{dc} \sum_{t \in T} \theta_{td} \theta_{tc},$$

где n_{dc} — вес связи между документами, например, число ссылок из d на c. В [59] предложена похожая модель LDA-JS, в которой вместо максимизации ковариации минимизируется дивергенция Йенсена-Шеннона между распределениями θ_d и θ_c. Формула М-шага для θ_{td}, согласно (18), приводит к следующей разновидности сглаживания:

$$\theta_{td} = \underset{t \in T}{\text{norm}}\Big(n_{td} + \tau \theta_{td} \sum_{c \in D} n_{dc} \theta_{tc}\Big).$$

Вероятности θ_{td} в ходе итераций приближаются к вероятностям θ_{tc} документов, связанных с d.

Регуляризатор матрицы Θ становится вычислительно неэффективным при пакетной обработке больших коллекций, когда документы c, на которые ссылается данный документ d, находятся в других пакетах. Проблема решается введением модальности документов, на которые есть ссылки из других документов. Этот способ порождает новую проблему: матрица Φ может не поместиться в оперативную память. Чтобы сократить модальность, можно оставить только документы c, число ссылок на которые $n_c = \sum_d n_{dc}$ больше заданного порога.

Данная идея пришла из модели влияния научных публикаций LDA-post [59]. В ней используются две модальности: слова W^1 и цитируемые документы $W^2 \subseteq D$. Модель выявляет наиболее влиятельные документы внутри каждой темы. Ненулевые элементы в строке c матрицы Φ_2 показывают, на какие темы повлиял документ $c \in W^2$. Также модель позволяет различать, какие из ссылок существенно повлияли на научную статью, а какие являются второстепенными. Считается, что документ c повлиял на документ d, если d ссылается на c и они имеют значительную долю общей тематики.

Геолокации. Информация о географическом положении часто используется при анализе данных социальных сетей. Географическая привязка документа d или его автора задаётся либо *геотегами* (названиями страны, региона, населённого пункта), либо *геолокацией* — парой географических координат $\ell_d = (x_d, y_d)$. В первом случае

вводится модальность геотегов, во втором используется регуляризатор. ARTM позволяет вводить в модель оба типа географических данных.

Целью моделирования может быть выделение региональных тем, определение «ареала обитания» каждой темы, поиск похожих тем в других регионах. Например, в качестве одной из иллюстраций в [206] определяются регионы популярности национальной кухни по постам пользователей Flickr. Другая иллюстрация из [117] показывает, что тематическая модель, учитывающая, из какого штата США пришло сообщение, точнее прослеживает путь урагана «Катрина».

Квадратичный регуляризатор матрицы Θ, предложенный в [206], формализует предположение, что документы со схожими геолокациями имеют схожую тематику:

$$R(\Theta) = -\frac{\tau}{2} \sum_{(c,d)} w_{cd} \sum_{t \in T} \big(\theta_{td} - \theta_{tc}\big)^2,$$

где w_{cd} — вес пары документов (c, d), выражающий близость геолокаций. Например, $w_{cd} = \exp(-\gamma r_{cd}^2)$, где $r_{cd}^2 = (x_c - x_d)^2 + (y_c - y_d)^2$ — квадрат евклидова расстояния.

Этот регуляризатор требует при обработке каждого документа d доступа к векторам θ_c других документов, что затрудняет пакетную обработку больших коллекций. Альтернативный способ сглаживания основан на регуляризации матрицы Φ.

Пусть G — модальность геотегов, $\varphi_{gt} = p(g\,|\,t)$. Тематика геотега g выражается по формуле Байеса: $p(t\,|\,g) = \varphi_{gt} \frac{n_t}{n_g}$, где n_g — частота геотега g в исходных данных, $n_t = \sum_g n_{gt}$ — частота темы t в модальности геотегов, вычисляемая EM-алгоритмом.

Квадратичный регуляризатор матрицы Φ по модальности геотегов формализует предположение, что географически близкие геотеги имеют схожую тематику:

$$R(\Phi) = -\frac{\tau}{2} \sum_{g,g' \in G} w_{gg'} \sum_{t \in T} n_t^2 \Big(\frac{\varphi_{gt}}{n_g} - \frac{\varphi_{g't}}{n_{g'}}\Big)^2,$$

где $w_{gg'}$ — вес пары геотегов (g, g'), выражающий их географическую близость. Ниже мы рассмотрим обобщение этого регуляризатора на более широкий класс задач.

Графы и социальные сети. В [117] предложена более общая тематическая модель NetPLSA, учитывающая произвольные графовые (сетевые) структуры на множестве документов. Пусть задан граф $\langle V, E\rangle$ с множеством вершин V и множеством рёбер E. Каждой его вершине $v \in V$ соответствует подмножество документов $D_v \subset D$. Например, в роли D_v может выступать отдельный документ, все статьи одного автора v, все посты из одного географического региона v, и т. д.

Тематика каждой вершины $v \in V$ выражается через параметры модели Θ:

$$p(t\,|\,v) = \sum_{d \in D_v} p(t\,|\,d)\, p(d\,|\,v) = \frac{1}{|D_v|} \sum_{d \in D_v} \theta_{td}.$$

В модели NetPLSA используется квадратичный регуляризатор:

$$R(\Theta) = -\frac{\tau}{2} \sum_{(u,v) \in E} w_{uv} \sum_{t \in T} \big(p(t\,|\,v) - p(t\,|\,u)\big)^2,$$

где веса w_{uv} рёбер графа (u, v) задаются естественным образом, когда в задаче есть соответствующая дополнительная информация. Например, если D_v — все статьи автора v, то в качестве веса ребра w_{uv} естественно взять число статей, написанных

авторами u и v в соавторстве. Если подобной информации нет, то вес полагается равным единице.

Этот регуляризатор требует при обработке каждого документа d доступа к векторам θ_c других документов, что затрудняет эффективную пакетную обработку больших коллекций. Альтернативный путь состоит в том, чтобы множество вершин графа V объявить модальностью и перейти к регуляризации матрицы Φ.

В каждый документ $d \in D_v$ добавим терм v модальности V. Выразим тематику вершины v через параметры Φ по формуле Байеса: $p(t\,|\,v) = p(v\,|\,t)\frac{p(t)}{p(v)} = \varphi_{vt}\frac{n_t}{|D_v|}$, где $n_t = \sum_v n_{vt}$ — частота темы t в модальности V, вычисляемая EM-алгоритмом.

Регуляризатор NetPLSA сохраняет прежний вид, но становится функцией от Φ:

$$R(\Phi) = -\frac{\tau}{2}\sum_{(u,v)\in E} w_{uv}\sum_{t\in T} n_t^2\Big(\frac{\varphi_{vt}}{|D_v|} - \frac{\varphi_{ut}}{|D_u|}\Big)^2. \tag{63}$$

Во многих приложениях важны направленности связей, которые квадратичный регуляризатор не учитывает. Например, связь (u, v) может означать ссылку из документа u на документ v. В модели iTopicModel [169] предполагается, что если $(u, v) \in E$, то тематика $p(t\,|\,u)$ шире тематики $p(t\,|\,v)$. Поэтому минимизируется сумма дивергенций $\mathrm{KL}\big(p(t\,|\,v)\,\|\,p(t\,|\,u)\big)$, причём условные распределения $p(t\,|\,v)$ можно выразить как через Θ, так и через Φ:

$$R(\Theta \text{ или } \Phi) = \frac{\tau}{2}\sum_{(u,v)\in E} w_{uv}\sum_{t\in T} p(t\,|\,v)\ln p(t\,|\,u).$$

Как показали эксперименты[3], регуляризация матрицы Φ приводит практически к тем же результатам, что и регуляризация Θ для моделей NetPLSA и iTopicModels.

Выводы по главе

- Данные о тематическом сходстве документов легко формализовать с помощью регуляризатора матрицы Θ.
- Менее очевидная формализация с помощью регуляризатора матрицы Φ приводит к сопоставимым результатам, но не затрудняет пакетную обработку больших коллекций.

[3] Булатов В. Г. Использование графовой структуры в тематическом моделировании. Магистерская диссертация, МФТИ, 2016.
`http://www.MachineLearning.ru/wiki/images/4/4d/Bulatov-2016-ms.pdf`

13 Моделирование иерархий и выбор числа тем

Существует ли в реальных текстовых коллекциях объективное, истинное или оптимальное число тем? Внутри любой темы возможны вариации или аспекты, которые можно рассматривать и как отдельные подтемы, и как частные проявления одной общей темы. Если тема представлена в недостаточном объёме и для её надёжного определения не хватает данных, то не присоединить ли её к ближайшей теме? Аналогичная проблема с неоднозначным выбором числа кластеров возникает и в задачах кластеризации, рис. 17.

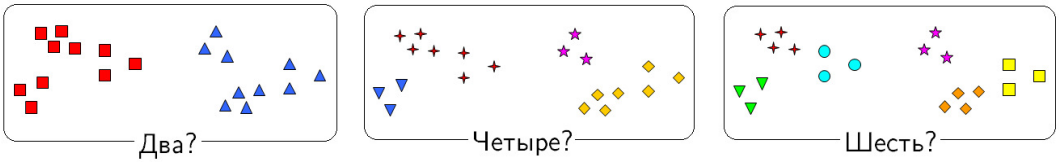


Рис. 17: Выбор числа кластеров в задачах кластеризации принципиально неоднозначен, как и выбор числа тем в задачах тематического моделирования.

Стоит ли разбивать темы на более мелкие подтемы, и как определить общее число тем — эти вопросы возникают в каждой практической задаче тематического моделирования. Зачастую выбор оказывается произвольным и субъективным. Использование статистических критериев для выявления значимых различий между темами не избавляет от субъективности, поскольку уровень значимости тоже выбирается эвристически. Вообще, многие известные методы, претендующие на объективность определения числа тем, содержат внутри себя параметр, от которого число тем явно или неявно зависит [180].

Эти соображения приводят к идее, что вместо поиска оптимального числа тем (которого может просто не существовать), имеет смысл строить иерархические тематические модели, в которых темы по необходимости дробятся на подтемы, а уровень детализации (*гранулированность*) тематического представления выбирается исходя из потребностей прикладной задачи.

Определение числа тем по внешним критериям используется в тех случаях, когда конечной целью (или одной из целей) тематического моделирования является решение задачи классификации [156], информационного поиска [81] или сегментации [153] текстов. Недостаточное число тем может приводить к снижению выразительных способностей модели и качества решения целевой задачи. Избыточное число тем может приводить к переобучению и снижению качества на независимых тестовых данных. Поэтому на практике число тем варьируют по заданной сетке значений, как правило весьма грубой, и определяют минимальное число тем, при котором задача решается с приемлемым качеством по одному или нескольким критериям. Типичным способом анализа является построение графика зависимости внешнего критерия от числа тем.

Энтропийное разреживание для отбора тем предложено в [180] для удаления незначимых тем из тематической модели. Идея заключается в разреживании рас-

пределения $p(t)$, которое выражается через параметры θ_{td}:

$$R(\Theta) = -\tau n \sum_{t\in T} \frac{1}{|T|} \ln p(t), \qquad p(t) = \sum_d p(d)\theta_{td}.$$

Подставим этот регуляризатор в формулу М-шага (18):

$$\theta_{td} = \underset{t\in T}{\operatorname{norm}}\Big(n_{td} - \tau \frac{n}{|T|}\frac{p(d)}{p(t)}\theta_{td}\Big).$$

Заменим θ_{td} в правой части равенства несмещённой оценкой $\frac{n_{td}}{n_d}$:

$$\theta_{td} = \underset{t\in T}{\operatorname{norm}}\Big(n_{td}\Big(1 - \tau\frac{n}{n_t|T|}\Big)\Big). \tag{64}$$

Этот регуляризатор разреживает целиком строки матрицы Θ. Если значение счётчика n_t в знаменателе достаточно мало, то все элементы t-й строки оказываются равными нулю, и тема t полностью исключается из модели. При использовании данного регуляризатора сначала устанавливается заведомо избыточное число тем $|T|$. В ходе итераций число нулевых строк матрицы Θ постепенное увеличивается.

Отбор тем в ARTM намного проще непараметрических байесовских моделей — иерархического процесса Дирихле (hierarchical Dirichlet process, HDP) [172] или процесса китайского ресторана (Chinese restaurant process, CRP) [38].

В обоих подходах, ARTM и HDP, имеется управляющий параметр, выбирая который, можно получать модели с числом тем, различающимся на порядки (в ARTM это коэффициент регуляризации τ, в HDP — коэффициент концентрации γ).

В [182] были проведены эксперименты на полусинтетических данных, представляющих собой смесь двух распределений $p(w\,|\,d)$ — реальной коллекции, для которой число тем неизвестно, и синтетической коллекции с заданным числом тем. Синтетическая коллекция строилась путём перемножения матриц $\Phi\Theta$, полученных в результате тематического моделирования той же реальной коллекции. Оказалось, что HDP и ARTM способны определять истинное число тем на синтетических и полусинтетических данных. ARTM определяет его точнее и устойчивее. Однако чем ближе полусинтетические данные к реальным, тем менее чётко различим диапазон значений гиперпараметров τ или γ, на котором восстанавливается правильное число тем. На реальных данных он неразличим вовсе, причём для обоих подходов. Таким образом, про оба подхода нельзя сказать, что они определяют оптимальное число тем.

По скорости вычислений `BigARTM` с регуляризатором отбора тем оказался в 100 раз быстрее свободно доступной реализации HDP.

В ходе экспериментов [182] также выяснилось, что регуляризатор отбора тем имеет полезный сопутствующий эффект: он удаляет из модели дублирующие, расщеплённые и линейно зависимые темы. Теоретическое обоснование этого эффекта остаётся открытой проблемой, доказать его пока не удалось.

Иерархическое тематическое моделирование. Иерархические тематические модели рекурсивно делят темы на подтемы. Тематические иерархии служат для построения рубрикаторов, систематизации больших объёмов текстовой информации, информационного поиска и навигации по большим мультидисциплинарным коллекциям. Задача автоматической рубрикации текстов сложна своей неоднозначностью

и субъективностью. Различия во мнениях экспертов относительно рубрикации документов могут достигать 40% [1]. Несмотря на обилие работ по иерархическим тематическим моделям [37, 104, 123, 207, 149, 191, 192, 193, 194], оптимизация размера и структуры иерархии остаётся открытой проблемой; более того, оценивание качества иерархий — также открытая проблема [207].

Стратегии построения тематических иерархий весьма разнообразны: нисходящие (дивизимные) и восходящие (агломеративные), представляющие иерархию деревом или многодольным графом, наращивающие граф по уровням или по вершинам, основанные на кластеризации документов или термов. Нельзя назвать какую-то из стратегий предпочтительной; у каждой есть свои достоинства и недостатки.

Вероятностная модель межуровневых связей. В [49] предложена нисходящая стратегия на основе ARTM. Иерархия представляется многодольным графом с фиксированным числом уровней и заданным числом тем на каждом уровне, возрастающим по уровням сверху вниз. Каждый уровень представляет собой обычную «плоскую» тематическую модель, поэтому время построения модели остаётся линейным по объёму коллекции.

Для моделирования связей между уровнями в модель вводятся параметры $\psi_{st} = p(s\,|\,t)$ — условные вероятности подтем s в темах t. В мультидисциплинарных коллекциях подтема может иметь несколько родительских тем. Например, «биоинформатика» имеет в качестве родительских «биологию» и «информатику», а «бустинг решающих деревьев» — «бустинг» и «решающие деревья». Поэтому представление иерархии многодольным графом предпочтительнее, чем деревом.

На верхнем уровне иерархии строится обычная плоская тематическая модель. Пусть модель ℓ-го уровня с множеством тем T уже построена, и требуется построить модель уровня $\ell+1$ с множеством дочерних тем S (subtopics) и бóльшим числом тем, $|S| > |T|$. Для родительских тем уже имеются частотные оценки $\hat{p}(w\,|\,t) = \frac{n_{wt}}{n_t}$, $t \in T$. Потребуем, чтобы они как можно лучше приближались вероятностными смесями дочерних тем, $p(w\,|\,t) \approx \sum_s p(w\,|\,s)\,p(s\,|\,t)$. Для этого будем максимизировать логарифм правдоподобия:

$$R(\Phi, \Psi) = \tau \sum_{t\in T} \sum_{w\in W} n_{wt} \ln \sum_{s\in S} \varphi_{ws}\psi_{st} \;\to\; \max_{\Phi,\Psi}, \tag{65}$$

где $\Psi = (\psi_{st})_{S\times T}$ — *матрица связей*, которая становится дополнительной матрицей параметров для тематической модели дочернего уровня.

Задача максимизации регуляризатора $R(\Phi, \Psi)$ с точностью до обозначений совпадает с основной задачей тематического моделирования (13), если считать родительские темы t *псевдодокументами* с частотами термов $n_{wt} = \tau n_t \varphi_{wt}$. Это означает, что дочерняя модель с бóльшим числом тем стремится как можно точнее описать не только исходные данные, но и всю совокупность родительских тем.

Вместо модификации формул М-шага данный регуляризатор можно реализовать совсем просто. Достаточно, построив родительский уровень из $|T|$ тем, добавить в коллекцию $|T|$ псевдодокументов с частотами термов n_{wt}. Матрица Ψ получится в столбцах матрицы Θ, соответствующих псевдодокументам, как показано на рис. 18.

Данный подход к моделированию тематических иерархий реализован в библиотеке `BigARTM`.

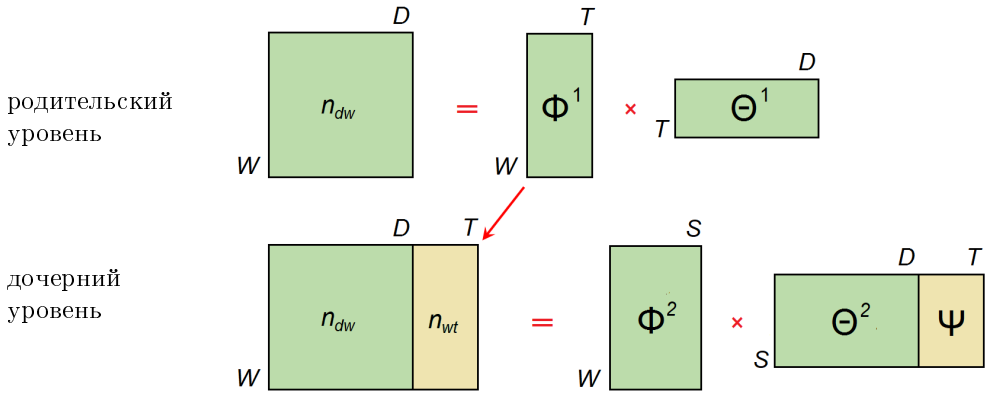


Рис. 18: Добавление второго уровня иерархии с множеством подтем S реализуется путём добавления в исходную коллекцию $|T|$ псевдодокументов с частотами термов n_{wt}. Матрица связей тем с подтемами $\Psi = \big(p(s\,|\,t)\big)$ образуется в столбцах матрицы Θ, соответствующих псевдодокументам.

Альтернативный подход, также предложенный в [49], заключается в том, чтобы приближать родительские темы как распределения $p(t\,|\,d)$ вероятностными смесями дочерних тем $\sum_{s\in S} p(t\,|\,s)p(s\,|\,d)$. В результате получается регуляризатор матрицы Θ для дочернего уровня, эквивалентный введению множества родительских тем T как модальности с частотами термов n_{td}. Эксперименты показали, что данный подход хуже приближает родительскую матрицу Φ^ℓ. Кроме того, добавление модальности в каждый документ труднее реализовать, чем добавление псевдодокументов в коллекцию. Данный подход был признан неудачным и не был реализован в `BigARTM`.

Разреживание межуровневых связей формализует естественное предположение, что каждая тема дочернего уровня $s \in S$ имеет небольшое число связей с темами родительского уровня $t \in T$. В частности, если все распределения $p(t\,|\,s)$ вырождены, то есть каждая тема s имеет только одну родительскую тему t, то вся иерархия приобретает вид дерева.

Применим кросс-энтропийный регуляризатор для разреживания распределений $p(t\,|\,s)$, выразив их через ψ_{st} по формуле Байеса:

$$R(\Psi) = -\tau \sum_{s\in S}\sum_{t\in T} \frac{1}{|T|} \ln p(t\,|\,s) = -\frac{\tau}{|T|} \sum_{t\in T}\sum_{s\in S} \ln \frac{\psi_{st} n_t}{\sum_z \psi_{sz} n_z}.$$

Поскольку матрица Ψ является частью матрицы Θ, к ней применима формула (18), из которой следует формула M-шага для модели дочернего уровня:

$$\psi_{st} = \underset{s\in S}{\operatorname{norm}}\Big(n_{st} + \tau\Big(p(t\,|\,s) - \frac{1}{|T|}\Big)\Big). \tag{66}$$

Согласно этой формуле, условные вероятности $p(t\,|\,s)$, меньшие $\frac{1}{|T|}$, становятся ещё меньше, и при достаточно большом τ обнуляются [49].

При разреживании распределений $p(s\,|\,t) = \frac{n_{st}}{n_t}$ важно, чтобы векторы $p(t\,|\,s) = \frac{n_{st}}{n_s}$ оставались распределениями, то есть чтобы у каждой подтемы s оставалась хотя бы одна родительская тема. На дочернем уровне S не должно образоваться ни одной *темы-сироты* s, которая совсем не имела бы родительских тем t.

1. остров, земля, период, там, территория, океан, где, более, вид, найти, вулкан, находиться, южный
2. растение, япония, раса, при, более, чем, например, исследование, вид, страна, население
3. вид, эволюция, самец, мозг, самка, животное, отбор, ген, более, птица, наш, между, чтобы, чем, друг
4. мозг, нейрон, при, заболевание, наш, пациент, состояние, система, болезнь, сон, исследование
5. клетка, музей, стволовой, ткань, организм, чтобы, опухоль, система, использовать, технология
6. клетка, ген, днк, организм, молекула, геном, белок, белка, бактерия, система, процесс, жизнь
7. система, материал, задача, структура, метод, компьютер, дать, при, химический, область, химия
8. квантовый, свет, волна, атом, информация, фотон, сигнал, использовать, два, при, частота, состояние
9. частица, энергия, кварк, взаимодействие, магнитный, электрон, масса, физика, бозон, протон, модель
10. звезда, галактика, земля, планета, вселенная, дыра, чёрный, объект, солнце, масса, наш, система
11. теория, пространство, вселенная, закон, физика, математический, уравнение, число, два, мир, система
12. наш, сеть, информация, дать, объект, культура, задача, например, образ, память, слово, разный
13. язык, слово, русский, например, говорить, словарь, речь, разный, языковой, текст, два, лингвист
14. наука, учёный, научный, потому, чтобы, лекция, хороший, университет, сейчас, наш, заниматься
15. экономический, экономика, страна, чтобы, более, рынок, компания, цена, решение, деньга, работа, чем
16. страна, война, государство, политический, россия, советский, власть, политика, германия, стать
17. ребёнок, женщина, мужчина, жизнь, культура, общество, себя, семья, социальный, советский, женский
18. город, пространство, социальный, городской, общество, место, культурный, жизнь, более, современный
19. исследование, социальный, поведение, группа, решение, and, the, теория, проблема, наука
20. социальный, социология, мир, теория, объект, социологический, действие, событие, социолог, наука
21. политический, философия, идея, наука, свобода, понятие, революция, история, философ, век, себя
22. право, власть, закон, король, век, римский, бог, себя, церковь, правовой, политический, суд, два
23. век, история, русский, исторический, имя, традиция, христианский, культура, историк, текст, уже
24. себя, искусство, литература, говорить, потому, мир, сам, миф, жизнь, слово, текст, роман, век
25. книга, фильм, автор, кино, pcourse, num, читатель, посвятить, тема, история, исследование, работа

Рис. 19: Пример плоского тематического спектра из 25 тем, построенного по коллекции публикаций научно-просветительского ресурса postnauka.ru на 2017 г.

Рис. 20: Пример иерархического тематического спектра из 10 и 40 тем, построенного по коллекции публикаций научно-просветительского ресурса postnauka.ru на 2017 г.

Спектр тем и визуализация иерархий. Тематические иерархии строятся в значительной степени для того, чтобы показывать их пользователям. До сих пор мы говорили о темах как о неупорядоченном множестве T. Однако просматривать список тем гораздо удобнее, если они ранжированы по семантической близости — так, чтобы близкие по смыслу темы оказывались в списке рядом. Это помогает пользователям быстрее находить нужные темы и акцентировать внимание на различиях между наиболее близкими темами. Это важно и для обычных «плоских» тематических моделей, и, тем более, для иерархических.

Тематическим спектром или *спектром тем* будем называть упорядоченный список тем $t_1, \ldots, t_{|T|}$, в котором сумма расстояний между соседними темами ми-

нимальна:

$$\sum_{i=2}^{|T|} \rho(t_i, t_{i-1}) \to \min,$$

где $\rho(t, t')$ — функция расстояния между темами как столбцами матрицы Φ. Например, может быть взято косинусное расстояние (97), расстояние Хеллингера (98) или расстояние Жаккара между семантическими ядрами тем (99).

Данная задача комбинаторной оптимизации эквивалентна задаче коммивояжёра[4]. Для её решения может быть использован, в частности, алгоритм Лина–Кернигана в реализации Хельсгауна [75], имеющий вычислительную сложность $O\big(|T|^{2.2}\big)$.

В тематическом спектре семантически близкие темы оказываются рядом, как показано на рис. 19. При построении иерархического спектра учитываются все уровни, чтобы минимизировать число пересечений в графе связей, см. рис. 20.

Выводы по главе

- Если тематическое моделирование является промежуточным этапом анализа данных, то выбирать число тем лучше всего по критерию качества решения конечной прикладной задачи.
- Иерархия даёт больше информации о тематической структуре коллекции, чем обычная плоская модель и позволяет выбирать необходимую степень детализации для каждой темы.
- Построение тематических иерархий в ARTM проще всего реализовать по уровням, сверху вниз. При построении дочернего уровня темы родительского уровня преобразуются в псевдо-документы, которые добавляются в коллекцию.
- На практике построение тематических иерархий затрудняется необходимостью подбирать число тем и коэффициенты регуляризации на каждом уровне.
- Построение спектра тем удобно для визуализации как плоских, так и иерархических тематических моделей.

[4] Федоряка Д. С. Технология интерактивной визуализации тематических моделей. Выпускная квалификационная работа бакалавра, МФТИ, 2017.
(`www.MachineLearning.ru/wiki/images/d/d8/Fedoriaka17bsc.pdf`).

14 Моделирование сочетаемости слов

Гипотеза «мешка слов» является одним из самых критикуемых предположений тематического моделирования. Она полностью игнорирует фундаментальное свойство *сочетаемости слов* (word co-occurrence) в естественном языке.

Сочетаемость бывает двух видов: *контактная* и *дистантная*.

Контактная сочетаемость означает частое появление некоторой последовательности из n подряд идущих слов, называемой *n-граммой*. Среди n-грамм представляют интерес коллокации и словосочетания. *Коллокация* — это n-грамма, встречающаяся в корпусе гораздо чаще, чем можно было бы ожидать, если бы слова генерировались случайно и независимо друг от друга. *Словосочетание* — это n-грамма, слова которой связаны грамматически и по смыслу, образуя единое понятие.

Тематические модели n-грамм (n-gram topic model) строятся на предположении, что все слова n-граммы порождаются одной и той же темой. Использование n-грамм, коллокаций или словосочетаний заметно улучшает интерпретируемость тем, что демонстрируется практически в каждой публикации по n-граммным тематическим моделям, см. например [85, 64]. Проблема в том, что число всех n-грамм катастрофически быстро растёт с ростом объёма коллекции. Поэтому выделяются не все подряд n-граммы, а только наиболее частотные.

Дистантная сочетаемость означает частое появление некоторой группы слов в одних и тех же контекстах, например, в одном предложении или в соседних предложениях. В отличие от *n-грамм*, слова не обязаны примыкать друг к другу.

Тематические модели дистантной сочетаемости опираются на гипотезу *дистрибутивной семантики*: «смысл слова определяется тем, в окружении каких слов оно чаще всего употребляется». Если два слова часто встречаются рядом, то, скорее всего, они порождаются одной и той же темой. Исходными данными для таких моделей служат не частоты слов в документах n_{dw}, а частоты пар слов n_{uv}, встречающихся в одних и тех же контекстах.

Существуют также тематические модели, обрабатывающие текст непосредственно как последовательность термов. Их мы рассмотрим в следующих главах 15 и 16.

Модели контактной сочетаемости. Первая биграммная тематическая модель BTM (bigram topic model) [188] представляла собой по сути мультимодальную модель, в которой каждому слову v соответствовала отдельная модальность со словарём $W^v \subseteq W$, составленным из всех слов, встречающихся непосредственно после слова v. Запишем log-правдоподобие этой модели в виде регуляризатора:

$$R(\Phi, \Theta) = \sum_{d\in D} \sum_{v\in d} \sum_{w\in W^v} n_{dvw} \ln \sum_{t\in T} \varphi^v_{wt}\theta_{td},$$

где $\varphi^v_{wt} = p(w\,|\,v,t)$ — условная вероятность слов w после слова v в теме t; n_{dvw} — частота биграммы «vw» в документе d. Главный недостаток модели BTM в том, что она учитывает только биграммы. Вторая проблема в том, что число всех биграмм быстро увеличивается с ростом коллекции, и использовать модель BTM на больших коллекциях затруднительно.

Модель TNG (topical n-grams) [197] устраняет эти недостатки. Условное распределение слов описывается вероятностной смесью

$$p(w\,|\,v,t) = \xi_{vwt}\varphi^v_{wt} + (1-\xi_{vwt})\varphi_{wt},$$

распознавание образов в биоинформатике		теория вычислительной сложности	
униграммы	**биграммы**	**униграммы**	**биграммы**
объект	задача распознавания	задача	разделять множества
задача	множество мотивов	множество	конечное множество
множество	система масок	подмножество	условие задачи
мотив	вторичная структура	условие	задача о покрытии
разрешимость	структура белка	класс	покрытие множества
выборка	распознавание вторичной	решение	сильный смысл
маска	состояние объекта	конечный	разделяющий комитет
распознавание	обучающая выборка	число	минимальный аффинный
информативность	оценка информативности	аффинный	аффинный комитет
состояние	множество объектов	случай	аффинный разделяющий
закономерность	разрешимость задачи	покрытие	общее положение
система	критерий разрешимости	общий	множество точек
структура	информативность мотива	пространство	случай задачи
значение	первичная структура	схема	общий случай
регулярность	тупиковое множество	комитет	задача MASC

Рис. 21: Примеры соответствующих тем в униграммной и биграммной модели (по коллекции статей всероссийской научной конференции «Математические методы распознавания образов»).

где ξ_{vwt} — переменная, равная вероятности того, что пара слов «vw» является биграммой в теме t. В работе С. С. Стенина[5] показано, что при некоторых не особо жёстких предположениях log-правдоподобие этой модели оценивается снизу взвешенной суммой log-правдоподобий модальностей униграмм и биграмм в модели ARTM. Другими словами, мультимодальная ARTM может быть использована для поиска приближённого решения в модели TNG.

В той же работе были проведены эксперименты с биграммной мультимодальной моделью ARTM на небольшой (менее 1000 документов) коллекции русскоязычных статей научной конференции ММРО (математические методы распознавания образов). Сопоставление тем униграммной и биграммной моделей показало, что по темам биграммной модели опрошенные постоянные участники конференции могли определить научную группу и даже авторов статей, тогда как по темам униграммной модели сделать это было проблематично, см. рис. 21.

В ARTM n-граммная модель естественным образом определяется как мультимодальная, в которой для каждого n выделяется отдельная модальность. Для предварительного сокращения словарей n-грамм подходит метод поиска коллокаций TopMine [64]. Он линейно масштабируется на большие коллекции и позволяет формировать словарь, в котором каждая n-грамма обладает тремя свойствами:

(а) имеет высокую частоту в коллекции;

(б) состоит из слов, неслучайно часто образующих n-грамму;

(в) не содержится в $(n{+}1)$-граммах, имеющих свойства (а) и (б).

Методы, предложенные в последующих работах, SegPhrase [107] и AutoPhrase [160], демонстрирующие ещё лучшие результаты.

Модель битермов. *Короткими текстами* (short text) называют документы, длина которых не достаточна для надёжного определения их тематики. Примерами коротких текстов являются сообщения Твиттера, заголовки новостных сообщений, рекламные объявления, реплики в диалогах, отдельные предложения, и т. д. Известны

[5] С. С. Стенин. Мультиграммные аддитивно регуляризованные тематические модели. Магистерская диссертация, ФИВТ МФТИ, 2015.
http://www.MachineLearning.ru/wiki/images/4/4a/Stenin2015MasterThesis.pdf

простые подходы к проблеме, но они не всегда применимы: объединять сообщения по какому-либо признаку (автору, времени, региону и т. д.); считать каждое сообщение отдельным документом, разреживая $p(t\,|\,d)$ вплоть до единственной темы; дополнять коллекцию длинными текстами (например, статьями Википедии). Одним из наиболее успешных и универсальных подходов к проблеме коротких текстов считается *тематическая модель битермов* (biterm topic model, BTM) [201].

Битермом называется пара слов, встречающихся рядом — в одном коротком сообщении или в одном предложении или в окне $\pm h$ слов. В отличие от биграммы, между двумя словами битерма могут находиться другие слова. Конкретизация понятия «рядом» зависит от постановки задачи и особенностей коллекции. Высокая частота битерма в текстовой коллекции является проявлением дистантной сочетаемости данной пары слов.

Модель BTM описывает вероятность появления пар слов (u, v). Исходными данными являются частоты n_{uv} битермов (u, v) в коллекции, или матрица вероятностей $P = (p_{uv})_{W\times W}$, где $p_{uv} = \underset{(u,v)\in W^2}{\operatorname{norm}}(n_{uv})$.

Примем гипотезу условной независимости $p(u, v\,|\,t) = p(u\,|\,t)\,p(v\,|\,t)$, то есть допустим, что слова u, v порождаются независимо друг от друга из одной и той же темы. Тогда, по формуле полной вероятности,

$$p(u, v) = \sum_{t\in T} p(u\,|\,t)\,p(v\,|\,t)\,p(t) = \sum_{t\in T} \varphi_{ut}\varphi_{vt}\pi_t,$$

где $\varphi_{wt} = p(w\,|\,t)$ и $\pi_t = p(t)$ — параметры тематической модели. Это трёхматричное разложение $P = \Phi\Pi\Phi^{\mathsf{T}}$, где $\Pi = \operatorname{diag}(\pi_1, \ldots, \pi_T)$ — диагональная матрица. Модель битермов не определяет тематику документов Θ и поэтому не подвержена влиянию эффектов, вызванных короткими текстами.

ARTM позволяет объединить модель битермов с обычной тематической моделью, чтобы всё-таки получить матрицу Θ. Для этого возьмём log-правдоподобие модели битермов в качестве регуляризатора с коэффициентом τ:

$$R(\Phi, \Pi) = \tau \sum_{u,v} n_{uv} \ln \sum_t \varphi_{ut}\varphi_{vt}\pi_t.$$

Применение уравнений (17)–(18) к этому регуляризатору даёт формулу M-шага для матрицы Φ:

$$\varphi_{wt} = \underset{w\in W}{\operatorname{norm}}\Big(n_{wt} + \tau \sum_{u\in W} n_{uw} p_{tuw}\Big); \tag{67}$$

$$p_{tuw} = \underset{t\in T}{\operatorname{norm}}\big(n_t \varphi_{wt}\varphi_{ut}\big). \tag{68}$$

Эти формулы можно интерпретировать как добавление в коллекцию *псевдо-документов* слов. Каждому слову $u \in W$ ставится в соответствие псевдо-документ d_u, объединяющий все контексты слова u. Это мешок слов, встретившихся рядом со словом u где-либо в коллекции. Число вхождений слова w в псевдо-документ d_u равно τn_{uw}. Вычисление вспомогательных переменных $p_{tuw} = p(t\,|\,u, w)$ в (68) в точности соответствует E-шагу для псевдо-документа d_u, если положить, что его тематический вектор образуется путём перенормировки строк матрицы Φ по формуле Байеса:

$$\theta_{tu} = \underset{t}{\operatorname{norm}}(n_t \varphi_{ut}). \tag{69}$$

Увеличивая коэффициент τ, можно добиться того, чтобы матрица Φ формировалась практически только по битермам. В таком случае модель ARTM переходит в модель битермов, которая строится по коллекции псевдо-документов, без использования исходных документов.

Модель сети слов. Идея моделировать не документы, а связи между словами, была положена в основу моделей WTM (word topic model) [47] и WNTM (word network topic model) [213]. Любопытно, что более ранняя публикация [47] осталась незамеченной (видимо, как не-байесовская), и статья [213] даже не ссылается на неё. Модели WTM и WNTM сводятся к применению PLSA и LDA соответственно к коллекции псевдо-документов d_u:

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

Запишем логарифм правдоподобия для тематической модели псевдо-документов $p(w\,|\,d_u)$ в виде регуляризатора:

$$R(\Phi,\Theta) = \tau \sum_{u,w\in W} n_{uw} \ln \sum_{t\in T} \varphi_{wt}\theta_{tu},$$

где n_{uw} — частота битерма (u,v) в коллекции (кстати, $n_{uw} = n_{wu}$).

Для объединения модели сети слов с обычной тематической моделью достаточно добавить псевдо-документы в основную коллекцию. В отличие от модели битермов, не придётся даже менять обычную формулу M-шага для вычисления θ_{tu}.

Если вычислять тематические векторы псевдо-документов по формуле (69), то модель WNTM переходит в модель битермов [146]. Более того, если в модели нет других регуляризаторов, то достаточно применить эту формулу только при инициализации матрицы Θ.

В экспериментах на коллекциях коротких текстов модель WNTM немного превосходит модель битермов и существенно превосходит обычные тематические модели [213]. На длинных документах тематические модели парной сочетаемости слов не показали значимых преимуществ перед обычными тематическими моделями.

Когерентность. Тема называется *когерентной* (согласованной), если наиболее частые термы данной темы часто встречаются рядом в документах коллекции [133]. Сочетаемость термов может оцениваться по самой коллекции D [125], или по сторонней коллекции, например, по Википедии [130]. Средняя когерентность тем считается хорошей мерой интерпретируемости тематической модели [134].

Пусть заданы оценки сочетаемости $C_{wv} = \hat{p}(w\,|\,v)$ для пар термов $(w,v) \in W^2$. Например, C_{wv} — доля документов, содержащих терм v, в которых терм w встречается не далее чем через 10 слов от v.

Запишем формулу полной вероятности, заменив условную вероятность φ_{vt} частотной оценкой:

$$\hat{p}(w\,|\,t) = \sum_{v\in W} \hat{p}(w\,|\,v)p(v\,|\,t) = \sum_{v\in W} C_{wv}\frac{n_{vt}}{n_t}.$$

Запишем в виде регуляризатора log-правдоподобия требование, чтобы параметры φ_{wt} приближали оценки $\hat{p}(w\,|\,t)$:

$$R(\Phi) = \tau \sum_{t\in T} n_t \sum_{w\in W} \hat{p}(w\,|\,t) \ln \varphi_{wt}.$$

Формула M-шага, согласно (17), принимает вид

$$\varphi_{wt} = \underset{w\in W}{\operatorname{norm}}\Big(n_{wt} + \tau \sum_{v\in W\setminus w} C_{wv} n_{vt}\Big). \tag{70}$$

Это сглаживающий регуляризатор. Он увеличивает вероятность терма в теме, если термы, с которыми он часто сочетается, относятся к данной теме. Точно такая же формула получилась в [125] для модели LDA и алгоритма сэмплирования Гиббса, но с более сложным обоснованием через обобщённую урновую схему Пойя, и с более сложной эвристической оценкой C_{wv}.

В работе [130] предложен другой регуляризатор когерентности:

$$R(\Phi) = \tau \sum_{t\in T} \ln \sum_{u,v\in W} C_{uv} \varphi_{ut} \varphi_{vt},$$

в котором оценка парной сочетаемости термов $C_{uv} = N_{uv}\big[\mathrm{PMI}(u,v) > 0\big]$ определяется через *поточечную взаимную информацию* (pointwise mutual information)

$$\mathrm{PMI}(u,v) = \ln \frac{|D| N_{uv}}{N_u N_v}, \tag{71}$$

где N_{uv} — число документов, в которых термы u, v хотя бы один раз встречаются рядом (не далее, чем через 10 слов), N_u — число документов, в которых терм u встречается хотя бы один раз.

Таким образом, единый подход к оптимизации когерентности пока не выработан. Предлагаемые критерии похожи на модели битермов и сети слов. Все они формализуют общую идею, что если слова часто совместно встречаются, то они имеют схожую тематику.

Модели векторных представлений слов ставят в соответствие каждому слову w вектор ν_w фиксированной размерности. Основное требование к этому отображению — чтобы близким по смыслу словам соответствовали близкие векторы. Согласно *дистрибутивной гипотезе* (distributional hypothesis) смысл слова определяется распределением слов, в окружении которых оно встречается [73]. Слова, встречающиеся в схожих контекстах, имеют схожую семантику и, соответственно, должны иметь близкие векторы. Для формализации этого принципа в [119, 120] предлагается несколько вероятностных моделей, все они реализованы в программе `word2vec`. В частности, модель skip-gram предсказывает появление слова w в контексте слова u, то есть при условии, что слово u находится рядом:

$$p(w\,|\,u) = \underset{w\in W}{\mathrm{SoftMax}}\,\langle \nu_w, \nu_u\rangle = \underset{w\in W}{\operatorname{norm}}\big(\exp\langle \nu_w, \nu_u\rangle\big) = \frac{\exp\langle \nu_w, \nu_u\rangle}{\sum_v \exp\langle \nu_v, \nu_u\rangle},$$

ассоциация	метод	ранжированный список слов
king – boy + girl	PWE	*queen*, princess, lord, prince
	W2V	*queen*, princess, regnant, kings
moscow – russia + spain	PWE	*madrid*, barcelona, aires, buenos
	W2V	*madrid*, barcelona, valladolid, malaga
india – russia + ruble	PWE	*rupee*, birbhum, pradesh, madhaya
	W2V	*rupee*, rupiah, devalued, debased
better – good + bad	PWE	really, something, thing, nothing
	W2V	*worse*, easier, prettier, funnier
cars – car + computer	PWE	*computers*, software, servers, implementations
	W2V	*computers*, software, hardware, microcomputers

Рис. 22: Сравнение списков ассоциаций, полученных моделями PWE и word2vec. Приводятся четыре наиболее близкие ассоциации.

где $\langle \nu_w, \nu_u \rangle = \sum_t \nu_{wt}\nu_{ut}$ — скалярное произведение векторов. В отличие от тематических моделей, нормировка вероятностей производится нелинейным преобразованием SoftMax, а сами векторные представления слов не нормируются.

Для обучения модели решается задача максимизации логарифма правдоподобия, как правило, градиентными методами:

$$\sum_{u,w \in W} n_{uw} \ln p(w \,|\, u) \;\to\; \max_{\{\nu_w\}}.$$

Постановка задачи очень похожа на обучение тематических моделей BTM и WNTM. Модели семейства `word2vec` и другие модели векторных представлений слов также являются матричными разложениями [102, 143, 108]. Главное отличие заключается в том, что в этих векторных представлениях координаты не интерпретируемы, не нормированы и не разрежены, тогда как в тематических моделях словам соответствуют разреженные дискретные распределения тем $p(t \,|\, w)$. С другой стороны, тематические модели изначально не предназначались для определения семантической близости слов, поэтому делают они это плохо.

В [146] предложен способ построения *тематических векторных представлений слов* (probabilistic word embedding, PWE) по псевдо-коллекции документов, аналогичный моделям BTM и WNTM. В задачах семантической близости слов они конкурируют с моделями `word2vec` и существенно превосходят обычные тематические модели. При этом тематические векторные представления являются интерпретируемыми и разреженными. Используя кросс-энтропийные регуляризаторы, разреженность векторов удаётся доводить до 93% без потери качества. На рис. 22 показаны примеры решения задачи ассоциаций слов с помощью моделей, построенных по англоязычной Википедии.

Количественные оценки показывают, что PWE решает задачи ассоциации слов намного лучше обычной тематической модели LDA, но не столь успешно конкурирует с лучшим моделям семейства `word2vec`, как в задачах семантической близости.

В задаче семантической близости документов PWE уверенно опережают векторную модель DBOW [54], специально разработанную для поиска семантически близких документов.

ARTM позволяет обобщить тематические модели дистрибутивной семантики для мультимодальных коллекций [146]. Используя данные о парной сочетаемости термов различных модальностей, возможно строить интерпретируемые тематические век-

торные представления для всех модальностей. В то же время, привлечение дополнительной информации о других модальностях повышает качество решения задачи близости слов.

Выводы по главе

- Учёт порядка слов в тексте и уход от гипотезы «мешка слов» — одно из важнейших направлений развития тематического моделирования.
- Простейший шаг в этом направлении — переход от «мешка слов» к «мешку словосочетаний». Одно это уже позволяет существенно улучшить интерпретируемость тем.
- Следующий шаг — переход от «мешка слов» к «сети слов». Использование данных о дистантной сочетаемости слов приводит к улучшению качества тематических векторных представлений слов и коротких текстов.
- Полноценный учёт порядка слов в тексте возможен только при его обработке как последовательности термов. К построению таких тематических моделей мы и перейдём в двух следующих главах.

15 Моделирование последовательного текста

Гипотеза «мешка слов» не настолько жёстко встроена в постановку задачи тематического моделирования, как может показаться. Представим текстовую коллекцию в виде последовательности $(d_i, w_i)_{i=1}^n$, где документы конкатенированы (записаны друг за другом), и в каждом документе сохранён естественный порядок слов. Запишем задачу максимизации логарифма правдоподобия:

$$\sum_{i=1}^{n} \ln \sum_{t \in T} p(w_i \,|\, t) p(t \,|\, d_i) = \sum_{i=1}^{n} \ln \sum_{t \in T} \varphi_{w_i t} \theta_{t d_i} \;\to\; \max_{\Phi, \Theta}.$$

Раньше мы приводили подобные слагаемые в этой сумме и записывали логарифм правдоподобия (13) через частоты термов в документах n_{dw}, теряя информацию о порядке слов внутри документов. Теперь мы обобщим модель, заменив документ d_i на контекст C_i — локальную окрестность терма w_i внутри документа d_i:

$$\sum_{i=1}^{n} \ln \sum_{t \in T} p(w_i \,|\, t) p(t \,|\, C_i) \;\to\; \max_{\Phi}.$$

Раньше контекстом терма был весь документ, $C_i \equiv d_i$. В данной главе мы ослабим это предположение, считая, что тематика терма w_i определяется не всем документом, а только его контекстом, например, предложением, абзацем или параграфом. В таком случае нам придётся не только задавать границы контекстов, но и моделировать тематику контекста $p(t \,|\, C_i)$ без введения матрицы Θ.

Это не только отказ от гипотезы «мешка слов», но также отказ от деления текстовой коллекции на документы. Границы документов и в самом деле довольно условны. Разбивать ли энциклопедию на статьи? Считать ли документом книгу, главу, параграф, абзац или тематически однородный сегмент? Теперь все эти вопросы решаются на этапе задания контекста C_i для каждого терма w_i. И, кстати, не следовало бы исключать терм w_i из своего собственного контекста C_i?

Однако начнём мы с промежуточного частного случая, когда контекст всё ещё совпадает с документом, но тематика контекста уже моделируется без матрицы Θ.

Однопроходный E-шаг. Вычисление тематического вектора документа $\theta_d = (\theta_{td})_{t \in T}$ в EM-алгоритме требует многих итераций по всем термам документа. В статье [11] предлагается вычислять вектор θ_d по явной формуле $\theta_{td} = \theta_{td}(\Phi)$ за один линейный проход документа d. Ограничение-равенство фактически выступает в роли регуляризатора, хотя и не записывается в виде оптимизационного критерия.

Данный подход имеет несколько преимуществ.

Во-первых, вектор θ_d вычисляется быстрее, не требуя итераций.

Во-вторых, исключение матрицы Θ из модели приводит к сокращению размерности и уменьшению переобучения. В обычной модели недостаточное качество матрицы Φ может быть скомпенсировано итерационной подгонкой каждого столбца θ_d матрицы Θ под конкретный документ d. Чем короче документ, тем менее надёжно обучение вектора θ_d, вплоть до заметного переобучения. Если же Θ выражается через Φ по фиксированной формуле, то переподгонка становится невозможной.

В-третьих, размер матрицы Θ линейно зависит от числа документов в коллекции, тогда как размер словаря увеличивается по сублинейному степенному закону Хипса [61]. Рост словаря может быть ограничен и принудительно, путём отбрасывания наименее частотных слов. Таким образом, возникают возможности для уменьшения темпа роста размерности модели при расширении коллекции.

Начнём с общего случая произвольной заданной зависимости $\theta_{td}(\Phi)$, затем рассмотрим частный случай линейной тематизации текста.

Теорема 15.1. *Пусть функции $\theta_{td}(\Phi)$ и $R(\Phi,\Theta)$ непрерывно дифференцируемы. Тогда точка Φ локального экстремума задачи (15) с ограничениями (14) и дополнительными 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:*

$$p_{tdw} = \underset{t\in T}{\operatorname{norm}}\big(\varphi_{wt}\theta_{td}\big); \tag{72}$$

$$n_{td} = \sum_{w\in d} n_{dw} p_{tdw} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}; \tag{73}$$

$$p'_{tdw} = p_{tdw} + \frac{\varphi_{wt}}{n_{dw}} \sum_{s\in T} \frac{n_{sd}}{\theta_{sd}} \frac{\partial \theta_{sd}}{\partial \varphi_{wt}}; \tag{74}$$

$$\varphi_{wt} = \underset{w\in W}{\operatorname{norm}}\left(\sum_{d\in D} n_{dw} p'_{tdw} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}}\right). \tag{75}$$

Доказательство. Воспользуемся необходимым условием экстремума задачи (15) с ограничениями (14), которые даёт следствие 4.2 из теоремы 4.1. Рассмотрим функционал, максимизируемый на М-шаге (19):

$$Q(\Phi) = \sum_{d\in D}\sum_{u\in W}\sum_{s\in T} n_{du} p_{sdu}\big(\ln\varphi_{us} + \ln\theta_{sd}(\Phi)\big) + R(\Phi,\Theta(\Phi)) \;\to\; \max_{\Phi}.$$

Запишем частные производные Q по параметрам φ_{wt}, выделяя в формулах выражения n_{sd} и p'_{tdw} согласно (73) и (74) соответственно:

$$\begin{aligned}
\varphi_{wt}\frac{\partial Q}{\partial \varphi_{wt}} &= \sum_{d\in D} n_{dw} p_{tdw} + \sum_{d,s,u} n_{du} p_{sdu} \frac{\varphi_{wt}}{\theta_{sd}} \frac{\partial \theta_{sd}}{\partial \varphi_{wt}} + \varphi_{wt} \sum_{d,s} \frac{\partial R}{\partial \theta_{sd}} \frac{\partial \theta_{sd}}{\partial \varphi_{wt}} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}} = \\
&= \sum_{d\in D} n_{dw}\left(p_{tdw} + \frac{\varphi_{wt}}{n_{dw}} \sum_{s\in T} \frac{1}{\theta_{sd}} \underbrace{\left(\sum_{u\in d} n_{du} p_{sdu} + \theta_{sd}\frac{\partial R}{\partial \theta_{sd}}\right)}_{n_{sd}} \frac{\partial \theta_{sd}}{\partial \varphi_{wt}}\right) + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}} = \\
&= \sum_{d\in D} n_{dw} \underbrace{\left(p_{tdw} + \frac{\varphi_{wt}}{n_{dw}} \sum_{s\in T} \frac{n_{sd}}{\theta_{sd}} \frac{\partial \theta_{sd}}{\partial \varphi_{wt}}\right)}_{p'_{tdw}} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}}.
\end{aligned}$$

Отсюда и из леммы 3.2 о максимизации на единичных симплексах следует формула М-шага (75).

Заметим, что θ_{sd} в знаменателе не может обращаться в нуль, поскольку при $\theta_{sd} = 0$ имеем $p_{sdu} = 0$ для всех u, следовательно, соответствующее слагаемое в сумме по s отсутствует (равно нулю) в функционале Q ещё до его дифференцирования.

Теорема доказана.

Таким образом, модификация EM-алгоритма сводится к *пост-обработке E-шага* — преобразованию переменных p_{tdw} в p'_{tdw}, которые затем подставляются в обычную формулу M-шага для матрицы Φ. С другой стороны, её можно рассматривать и как аддитивную поправку в формуле M-шага, то есть ограничение-равенство $\theta_{td} = \theta_{td}(\Phi)$ действительно играет роль регуляризатора.

Теперь настало время распорядиться свободой выбора функций $\theta_{td}(\Phi)$.

Линейная тематизация текста. Для оценивания тематического вектора документа $p(t\,|\,d)$ запишем формулу полной вероятности, заменив в ней распределения $p(t\,|\,w,d)$ на распределения $p(t\,|\,w)$, а для получения тематических векторов термов $p(t\,|\,w)$ из столбцов матрицы Φ применим формулу Байеса:

$$\theta_{td}(\Phi) = \sum_{w\in d} p(t\,|\,d,w)\,p(w\,|\,d) = \sum_{w\in d} p(w\,|\,d)\,p(t\,|\,w) = \sum_{w\in d} p_{wd} \operatorname*{norm}_{t\in T}\big(\varphi_{wt}p_t\big), \tag{76}$$

где $p_{wd} = \frac{n_{dw}}{n_d}$ — частотная оценка распределения $p(w\,|\,d)$, $p_t = \frac{n_t}{n}$ — частотная оценка распределения $p(t)$, без труда получаемая из EM-алгоритма.

Будем называть уравнение (76) *линейной тематизацией текста*, так как тематический вектор документа $p(t\,|\,d)$ вычисляется путём усреднения тематических векторов термов $p(t\,|\,w)$ за один линейный проход по документу.

Альтернативное обоснование линейной тематизации можно получить, записав тематический вектор документа $p(t\,|\,d)$ после первой итерации EM-алгоритма при начальном приближении $\theta^0_{td} = p_t$:

$$\theta_{td}(\Phi) = \operatorname*{norm}_{t\in T}\Bigg(\sum_{w\in W} n_{dw}p_{tdw}\Bigg) = \sum_{w\in d}\frac{n_{dw}}{n_d}\frac{\varphi_{wt}\theta^0_{td}}{\sum_s \varphi_{ws}\theta^0_{sd}} = \sum_{w\in d} p_{wd}\operatorname*{norm}_{t\in T}\big(\varphi_{wt}p_t\big).$$

Применим теорему 15.1 к формуле (76). Для этого найдём частную производную:

$$\begin{aligned}\varphi_{wt}\frac{\partial\theta_{sd}}{\partial\varphi_{wt}} &= \varphi_{wt}\frac{\partial}{\partial\varphi_{wt}}\left(\frac{p_{wd}\varphi_{ws}p_s}{\sum_v \varphi_{wv}p_v}\right) = \\ &= \varphi_{wt}p_t p_{wd}\frac{\delta_{st}\sum_v \varphi_{wv}p_v - \varphi_{ws}p_s}{\big(\sum_v \varphi_{wv}p_v\big)^2} = p_{wd}\varphi'_{tw}\big(\delta_{st} - \varphi'_{sw}\big),\end{aligned}$$

где $\delta_{st} = [s{=}t]$ — символ Кронекера, $\varphi'_{tw} = \frac{\varphi_{wt}p_t}{\sum_v \varphi_{wv}p_v} = \operatorname{norm}_t(\varphi_{wt}p_t) = p(t\,|\,w)$ — результат перенормировки строки w матрицы Φ по формуле Байеса. Подставим это выражение в (75) и перепишем уравнения в порядке, удобном для проведения вы-

числений методом простых итераций [11]:

$$\begin{aligned}
\varphi'_{tw} &= \underset{t\in T}{\operatorname{norm}}\big(\varphi_{wt} n_t\big);\\
\theta_{td} &= \sum_{w\in d} p_{wd}\varphi'_{tw};\\
p_{tdw} &= \underset{t\in T}{\operatorname{norm}}\big(\varphi_{wt}\theta_{td}\big);\\
n_t &= \sum_{d\in D}\sum_{w\in d} n_{dw} p_{tdw};\\
n_{td} &= \sum_{w\in d} n_{dw} p_{tdw} + \theta_{td}\frac{\partial R}{\partial \theta_{td}};\\
p'_{tdw} &= p_{tdw} + \frac{\varphi'_{tw}}{n_d}\left(\frac{n_{td}}{\theta_{td}} - \sum_{s\in T}\varphi'_{sw}\frac{n_{sd}}{\theta_{sd}}\right);\\
\varphi_{wt} &= \underset{w\in W}{\operatorname{norm}}\left(\sum_{d\in D} n_{dw} p'_{tdw} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}}\right).
\end{aligned} \tag{77}$$

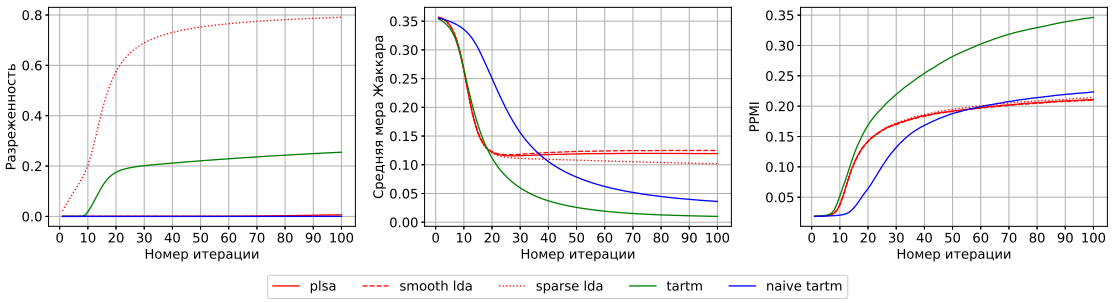
Вычисление переменных φ'_{tw}, θ_{td}, p_{tdw}, n_t, n_{td}, p'_{tdw} образует E-шаг и занимает $O(n_d|T|)$ операций по каждому документу, как в обычном EM-алгоритме. Все переменные, относящиеся к документу d, можно удалять из памяти по окончании его обработки. Вычисление переменных φ_{wt} на M-шаге происходит по обычной формуле, только вместо условных вероятностей p_{tdw} подставляются переменные p'_{tdw}. Таким образом, модификация EM-алгоритма не приводит к существенному увеличению времени его работы или дополнительному расходу памяти.

Обработка каждого документа на E-шаге осуществляется за два прохода. На первом проходе вычисляются переменные φ'_{tw} и θ_{td}. На втором проходе вычисляются остальные переменные p_{tdw}, n_t, n_{td}, p'_{tdw}. Если требуется только найти тематический вектор документа θ_d, не обновляя матрицу Φ, то второй проход делать не нужно. Нормировочные множители для строк матрицы Φ можно вычислять и сохранять после каждого обновления матрицы Φ.

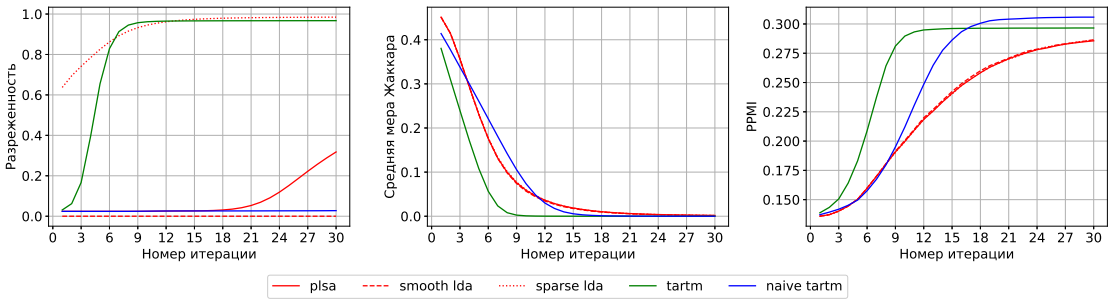
Вычисления можно ещё упростить, если отказаться от регуляризации по Θ. Тогда $\frac{\partial R}{\partial \theta_{td}} = 0$. Если теперь воспользоваться правилом подстановки несмещённых оценок максимального правдоподобия (21), то после подстановки $\frac{n_{td}}{\theta_{td}} = \frac{n_{sd}}{\theta_{sd}} = n_d$ в (77) окажется, что $p'_{tdw} = p_{tdw}$. Тогда вычислять n_{td} и p'_{tdw} вообще не нужно:

$$\begin{aligned}
\varphi'_{tw} &= \underset{t\in T}{\operatorname{norm}}\big(\varphi_{wt} n_t\big);\\
\theta_{td} &= \sum_{w\in d} p_{wd}\varphi'_{tw};\\
p_{tdw} &= \underset{t\in T}{\operatorname{norm}}\big(\varphi_{wt}\theta_{td}\big);\\
n_t &= \sum_{d\in D}\sum_{w\in d} n_{dw} p_{tdw};\\
\varphi_{wt} &= \underset{w\in W}{\operatorname{norm}}\left(\sum_{d\in D} n_{dw} p_{tdw} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}}\right).
\end{aligned}$$

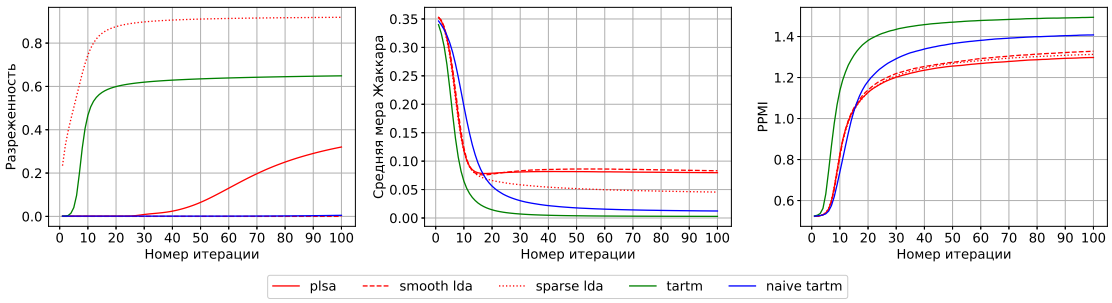
Этот вариант EM-алгоритма отличается от обычного простым быстрым вычислением тематических векторов $p(t\,|\,w)$ и $p(t\,|\,d) = \theta_{td}$ за один проход по документу.



(а) коллекция NIPS, число тем $|T| = 50$



(б) коллекция Twitter, число тем $|T| = 50$



(в) коллекция 20-newsgroups, число тем $|T| = 25$

Рис. 23: Три критерия качества (разреженность матрицы Φ, средняя близость тем по мере Жаккара, средняя когерентность тем по PPMI) для пяти моделей (PLSA, LDA со сглаживанием, LDA с разреживанием, TARTM, «наивный» TARTM) по трём текстовым коллекциям.

Эксперименты с линейной тематизацией показали, что она одновременно улучшает несколько критериев качества, даже без дополнительных регуляризаторов [11]. Повышается скорость сходимости, разреженность, различность и когерентность тем, при этом темы очищаются от общеупотребительных слов.

Для сравнения с `TARTM` (Θ-less ARTM) строилась также модель `naive TARTM` — обычный EM-алгоритм с одной итерацией по документу, чтобы показать, что линейная тематизация не сводится к такому наивному упрощению, см. рис. 23.

Локализованный E-шаг. Формула $\theta_{td}(\Phi)$ определяет тематический вектор $p(t\,|\,d)$ документа $d = (w_1, \ldots, w_{n_d})$ путём усреднения векторов $p(t\,|\,w) = \varphi'_{tw}$ всех его термов:

$$\theta_{td}(\Phi) \equiv p(t\,|\,d) = \frac{1}{n_d}\sum_{i=1}^{n_d} p(t\,|\,w_i) = \frac{1}{n_d}\sum_{i=1}^{n_d} \varphi'_{tw_i}.$$

Аналогичным образом можно определить тематический вектор для произвольного текстового фрагмента — словосочетания, предложения, абзаца, раздела.

Локальным контекстом терма w_i в документе d будем называть подпоследовательность термов $C_i = \{w_b, \ldots, w_e\}$, находящихся в окрестности терма w_i. Мы оставляем большую свободу в выборе контекста: он может быть левым, правым или двусторонним; также может пропускать некоторые термы, в том числе и сам w_i.

Тематический вектор локального контекста C_i определяется путём усреднения векторов $p(t\,|\,u) = \varphi'_{tu}$ всех его термов:

$$\theta_{ti}(\Phi) \equiv p(t\,|\,i) = \frac{1}{|C_i|}\sum_{u\in C_i} p(t\,|\,u) = \frac{1}{|C_i|}\sum_{u\in C_i} \varphi'_{tu}.$$

Некоторые термы контекста C_i могут быть более важны для определения его тематики. Поэтому сразу обобщим определение и будем усреднять тематические векторы $p(t\,|\,u)$ с неотрицательными нормированными весами $\alpha(u\,|\,i)$:

$$\theta_{ti}(\Phi) \equiv p(t\,|\,i) = \sum_{u\in C_i} \varphi'_{tu}\alpha(u\,|\,i), \qquad \sum_{u\in C_i} \alpha(u\,|\,i) = 1, \qquad \alpha(u\,|\,i) \geqslant 0. \tag{78}$$

Замена тематики документа $\theta_{td} = p(t\,|\,d)$ тематикой локального контекста $\theta_{ti} = p(t\,|\,i)$ в i-й позиции приводит к конструкции *локализованной тематической модели*, в которой для предсказания терма используется не весь документ, а только ближайшее окружение данного терма, его контекст:

$$p(w\,|\,d,i) = \sum_{t\in T} p(w\,|\,t)\,p(t\,|\,i) = \sum_{t\in T} \varphi_{wt} \sum_{u\in C_i} \varphi'_{tu}\alpha(u\,|\,i).$$

Локализованная тематическая модель переходит в обычную, если в качестве контекста C_i взять весь документ, а веса распределить равномерно: $\alpha(u\,|\,i) = \frac{1}{n_d}$.

Локализация ослабляет исходные допущения тематического моделирования. Это не только отказ от гипотезы «мешка слов», но также отказ от деления текстовой коллекции на документы.

Запишем EM-алгоритм для локализованной тематической модели в общем виде. Будем полагать, что термы в коллекции имеют сквозную нумерацию $i = 1, \ldots, n$ и что для каждой позиции i определён контекст C_i и распределение весов $\alpha(u\,|\,i)$.

$$\begin{aligned}
\varphi'_{tw} &= \operatorname*{norm}_{t\in T}\big(\varphi_{wt}n_t\big);\\
\theta_{ti} &= \sum_{u\in C_i} \varphi'_{tu}\alpha(u\,|\,i);\\
p_{ti} &= \operatorname*{norm}_{t\in T}\big(\varphi_{w_it}\theta_{ti}\big);\\
n_t &= \sum_{i=1}^{n} p_{ti};\\
\varphi_{wt} &= \operatorname*{norm}_{w\in W}\bigg(\sum_{i=1}^{n}[w_i = w]p_{ti} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}}\bigg).
\end{aligned}$$

Множество документов D и отдельные документы d в этом алгоритме вообще не фигурируют. Тематика любого документа, как и любого текстового фрагмента, вычисляется «на лету» путём усреднения тематических векторов всех его термов.

Локализованный E-шаг с экспоненциальными скользящими средними. Введение весовых коэффициентов $\alpha(u\,|\,i)$ позволяет сделать границы контекстов нечёткими, уменьшая веса по мере увеличения дистанции между термами w_i и u.

В частности, задать веса можно с помощью двух *экспоненциальных скользящих средних* (ЭСС), которые усредняют векторы термов по рекуррентным формулам, пробегая по тексту в двух направлениях — слева направо и справа налево:

$$\vec{p}(t\,|\,i) = \gamma_i \cdot \varphi'_{tw_i} + (1-\gamma_i)\cdot \vec{p}(t\,|\,i-1), \quad i = 1,\ldots,n_d, \quad \gamma_1 = 1;$$
$$\overleftarrow{p}(t\,|\,i) = \gamma_i \cdot \varphi'_{tw_i} + (1-\gamma_i)\cdot \overleftarrow{p}(t\,|\,i+1), \quad i = n_d,\ldots,1, \quad \gamma_{n_d} = 1.$$

Для выбора коэффициента сглаживания $\gamma_i \in [0,1]$ можно ориентироваться на известную для ЭСС оценку $\gamma_i \approx \frac{1}{h}$, где h — число усредняемых векторов. Например, $\gamma_i = 0.1$ приблизительно соответствует усреднению 10 векторов.

Если $\gamma_i = \gamma$ — константа, то экспоненциальное скользящее среднее определяет веса $\alpha(w_k\,|\,i) = \gamma(1-\gamma)^{|i-k|}$.

Коэффициент γ_i можно менять в зависимости от позиции i. Его увеличение вплоть до $\gamma_i = 1$ позволяет «забыть» накопленную к данному моменту информацию о контексте, например, при переходе к следующему документу или к следующей секции документа. Уменьшение γ_i вплоть до нуля позволяет игнорировать терм, например, если это стоп-слово или слово общей лексики.

Зная тематические векторы левого и правого контекста $\vec{p}(t\,|\,i)$ и $\overleftarrow{p}(t\,|\,i)$ для каждой позиции i, можно отвечать на многие вопросы о тематической структуре текста и любых его фрагментов.

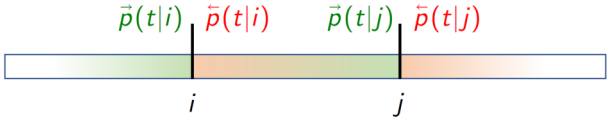
Тематика двустороннего контекста терма w_i оценивается как среднее арифметическое или (для общности) среднее взвешенное левого и правого контекста:

$$\theta_{ti}(\Phi) \equiv p(t\,|\,i) = \beta\vec{p}(t\,|\,i) + (1-\beta)\overleftarrow{p}(t\,|\,i).$$

Границу i между тематическими сегментами длинного текстового документа можно определять по максимуму различия между распределениями $\vec{p}(t\,|\,i)$ и $\overleftarrow{p}(t\,|\,i)$.

Анализ отклонений коротких и длинных ЭСС, соответственно при больших и малых γ_i, позволяет определить, содержит ли текст вкрапления тематически разнородных фрагментов, и где они находятся.

Рассмотрим *двунаправленные тематические векторы* для пары термов w_i и w_j:



Тематику короткого сегмента $[i\ldots j]$ можно оценивать как среднее арифметическое правого контекста его начала i и левого контекста его конца j:

$$p(t\,|\,i\ldots j) = \tfrac{1}{2}\big(\overleftarrow{p}(t\,|\,i) + \vec{p}(t\,|\,j)\big).$$

Тематическая однородность сегмента $[i\ldots j]$ оценивается тем, насколько схожи распределения $\overleftarrow{p}(t\,|\,i)$ и $\vec{p}(t\,|\,j)$.

Модели внимания и их связь с локализованным E-шагом. Модель внимания является основным «строительным блоком» современных нейросетевых моделей языка [29, 58, 98]. Она определяет вектор контекста y_i в позиции i как выпуклую комбинацию векторов x_u всех термов из контекста C_i:

$$y_i = \sum_{u \in C_i} \alpha(u\,|\,i)\, x_u, \qquad \sum_{u \in C_i} \alpha(u\,|\,i) = 1, \qquad \alpha(u\,|\,i) \geqslant 0. \tag{79}$$

Таким образом, смысл терма w_i, закодированный вектором x_i, обогащается смыслами окружающих его термов $u \in C_i$, образуя в результате вектор контекста y_i.

По сути, это та же формула (78), только нейросетевые векторные представления (эмбединги) термов не обязаны быть неотрицательными и нормированными.

Более важное отличие в том, что здесь коэффициенты $\alpha(u\,|\,i)$ оценивают не позиционную, а семантическую близость термов w_i и u через скалярное произведение их векторов $\langle x_i, x_u \rangle$. С бóльшим весом учитываются термы u, которые по смыслу близки к терму w_i, то есть векторы которых более похожи на вектор x_i:

$$\alpha(u\,|\,i) = \operatorname*{norm}_{u \in C_i} \exp\langle x_i, x_u \rangle.$$

Третье существенное отличие в том, что нейросетевая модель внимания дополнительно параметризуется тремя матрицами Q, K, V для линейного преобразования векторов термов, когда они выступают в трёх разных ролях:

$$y_i = \sum_{u \in C_i} (Vx_u) \operatorname*{norm}_{u \in C_i} \exp\langle Qx_i, Kx_u \rangle. \tag{80}$$

Модель внимания трансформирует исходный бесконтекстный вектор x_i терма w_i в контекстный вектор y_i, кодирующий ещё больше информации о смысле этого терма с учётом его локального контекста. Вектор Qx_i называется *запросом* (query), вектор Kx_u — *ключом* (key), вектор Vx_u — *значением* (value). Матрицы параметров Q, K, V обучаются по данным, что делает модель внимания намного более гибкой.

Формула внимания (80) описывает трёхслойную нейронную сеть. *Трансформером* называют нейросетевую архитектуру, которая выполняет несколько таких преобразований друг за другом — когда последовательность (y_i) с выхода предыдущей модели внимания подаётся в качестве последовательности (x_i) на вход последующей модели внимания. Каждый блок трансформера, помимо модели внимания, содержит, как правило, ещё два полносвязных слоя и два слоя нормировки, итого семь слоёв с огромным количеством обучаемых параметров. В одной из первых моделей машинного перевода [29] использовалось шесть подряд идущих блоков трансформера, вместе образующих глубокую нейронную сеть из 42 слоёв.

Несмотря на перечисленные выше различия, аналогия локализованного E-шага с моделью внимания несколько более глубокая. Аналогом вектора контекста y_i в (80) является не тематический вектор локального контекста $p(t\,|\,i) = \theta_{ti}$, а тематический вектор терма w_i в контексте C_i:

$$p(t\,|\,C_i, w_i) = p_{ti} = \operatorname*{norm}_{t \in T}\big(\varphi_{w_i t}\theta_{ti}\big) = \operatorname*{norm}_{t \in T}\Bigg(\sum_{u \in C_i} \varphi'_{tu}\varphi_{w_i t}\,\alpha(u\,|\,i)\Bigg).$$

Здесь тематические векторы $\big(\varphi'_{tu} = p(t\,|\,u)\big)_{t \in T}$ не просто складываются с весами, но сначала домножаются покомпонентно на вектор $\big(\varphi_{w_i t} = p(w_i\,|\,t)\big)_{t \in T}$. Он играет

Алгоритм 5. EM-алгоритм с локализованным E-шагом.

Вход: коллекция, число тем $|T|$, параметры K, L, β, $\vec{\gamma}_i$, $\overleftarrow{\gamma}_i$;
Выход: матрица Φ, векторы термов документов p_{ti};

1 инициализация $\varphi_{wt} := \mathrm{norm}_w(\mathrm{rand})$; $n_t := 1$ для всех $w \in W$, $t \in T$;
2 **для всех** итераций $k = 1, \ldots, K$ (проходов по всей коллекции)
3 инициализация: $n_{wt} := 0$; $\tilde{n}_{wt} := 0$; $\tilde{n}_t := 0$; для всех $w \in W$, $t \in T$;
4 **для всех** документов $d \in D$
5 $p_{ti} := \mathrm{norm}_t(\varphi_{w_i t} n_t)$ для всех $t \in T$, $i = 1, \ldots, n_d$;
6 **для всех** $l = 1, \ldots, L$ (аналог L блоков внимания)
7 $\vec{\theta}_{ti} := \vec{\gamma}_i\, p_{ti} + (1-\vec{\gamma}_i)\, \vec{\theta}_{t,i-1}$ для всех $t \in T$, $i = 1, \ldots, n_d$, $\vec{\gamma}_1 = 1$;
8 $\overleftarrow{\theta}_{ti} := \overleftarrow{\gamma}_i\, p_{ti} + (1-\overleftarrow{\gamma}_i)\, \overleftarrow{\theta}_{t,i+1}$ для всех $t \in T$, $i = n_d, \ldots, 1$, $\overleftarrow{\gamma}_{n_d} = 1$;
9 $p_{ti} := \mathrm{norm}_t\big(\big(\beta\vec{\theta}_{ti} + (1-\beta)\overleftarrow{\theta}_{ti}\big)p_{ti}/n_t\big)$ для всех $t \in T$, $i = 1, \ldots, n_d$;
10 $\tilde{n}_{w_i t} := \tilde{n}_{w_i t} + p_{ti}$; $\tilde{n}_t := \tilde{n}_t + p_{ti}$ для всех $t \in T$, $i = 1, \ldots, n_d$;
11 **если** пора обновить матрицу Φ **то**
12 $n_{wt} := n_{wt} + \tilde{n}_{wt}$; $\tilde{n}_{wt} := 0$ для всех $t \in T$, $w \in W$;
13 $\varphi_{wt} := \underset{w \in W}{\mathrm{norm}}\Big(n_{wt} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}}\Big)$ для всех $t \in T$, $w \in W$;
14 $n_t := \tilde{n}_t$ для всех $t \in T$;

роль фильтра, пропуская только те темы, которые есть у терма w_i. В результате сумма формируется преимущественно из векторов термов u, тематически схожих с термом w_i. Похожий эффект достигается и в модели внимания Query–Key–Value, когда векторы термов x_u домножаются на веса $\alpha(u\,|\,i)$, пропорциональные скалярным произведениям $\langle x_i, x_u \rangle$. Локализованный E-шаг эксплуатирует альтернативный механизм — покоординатную фильтрацию усредняемых векторов.

Таким образом, локализованный E-шаг можно рассматривать как предельно упрощённую модель внимания без обучаемых параметров. Он трансформирует последовательность бесконтекстных тематических векторов $p(t\,|\,w_i)$ в последовательность контекстных векторов $p(t\,|\,C_i, w_i)$.

EM-алгоритм с локализованным E-шагом. По аналогии с трансформером, полученные контекстные векторы $p(t\,|\,C_i, w_i)$ можно повторно подавать на вход локализованного E-шага вместо бесконтекстных векторов $p(t\,|\,w_i)$. В трансформере это привело бы к наращиванию архитектуры сети ещё семью слоями с обучаемыми параметрами. Однако в нашей (предельно упрощённой) тематической модели внимания никаких обучаемых параметров трансформация не имеет. Это просто ещё одна итерация по документу, см. алгоритм 5.

На входе алгоритма задаётся число итераций по всей коллекции K и число итераций по каждому документу L. Перед первым проходом документа (шаг 5) в массив p_{ti} записываются бесконтекстные векторы термов, $p_{ti} = p(t\,|\,w_i)$. Они обрабатываются по формулам локализованного E-шага (шаги 7–9). В конце прохода (шаг 9) в этом же массиве оказываются контекстные векторы термов $p_{ti} = p(t\,|\,C_i, w_i)$, которые на следующем цикле по l обрабатываются вместо бесконтекстных.

Обновления матрицы Φ (шаги 11–13) производятся по мере накопления достаточного объёма статистики в счётчиках n_{wt} — не слишком редко, но и не слишком часто, в зависимости от размеров коллекции.

Экспериментальное подтверждение работоспособности EM-алгоритма с локализованным E-шагом оставалось открытой задачей на момент написания данной главы.

Выводы по главе

- Линейная тематизация документов оказалась успешной стратегией в преодолении ограничений гипотезы «мешка слов». Определённо, это заявка на роль нового стандарта де-факто в тематическом моделировании.
- Локализованный E-шаг является предельно упрощённым аналогом модели внимания. Это новая модель, требующая экспериментальной проверки, а также выработки рекомендаций по подбору параметров L, β, $\vec{\gamma}_i$, $\check{\gamma}_i$.
- Остаётся открытым вопрос, возможна ли дальнейшая гибридизация локализованных тематических моделей и нейросетевых моделей внимания.
- Остаётся открытым вопрос, насколько обоснованы трансформации контекстных векторов $p(t\,|\,C_i, w_i)$, и как меняется при этом постановка исходной оптимизационной задачи. Частичный ответ на него будет дан в следующей главе.

16 Моделирование сегментированного текста

Гипотеза «мешка слов» и предположение о статистической независимости соседних слов приводят к слишком частой хаотичной смене тематики между соседними словами. Если проследить, к каким темам относятся последовательные слова в тексте, то сочетания тем внутри каждого предложения могут оказаться плохо интерпретируемыми, даже при том, что сами темы интерпретируются хорошо по ранжированным спискам частых слов.

В данной главе рассматриваются тематические модели, позволяющие учитывать более реалистичные предположения о связности текста, происходящие из теории дискурса и коммуникативной лингвистики [4]. Приведём лишь несколько примеров:

- каждое предложение, как правило, несёт сообщение о взаимодействии двух агентов, и потому относится к одной или двум темам;
- следующее предложение, как правило, продолжает тематику предыдущего;
- смена тематики чаще происходит между абзацами, чем между предложениями, ещё чаще — между секциями документа.

Начнём с тематических моделей, основанных на гипотезе, что каждое предложение является «мешком термов» и порождается одной темой. Документ при этом считается «мешком предложений». Затем введём механизм регуляризации E-шага, когда ограничения накладываются на последовательность векторов $p(t\,|\,d, w_i)$.

Тематическая модель предложений. Допустим, что каждый документ d разбит на множество сегментов S_d. Это могут быть предложения, абзацы или *фразы* — синтаксически корректные части предложений. Обозначим через n_s длину сегмента s, через n_{sw} — число вхождений слова w в сегмент s.

Предположим, что все слова сегмента относятся к одной теме и запишем функцию вероятности сегмента $s \in S_d$ через параметры тематической модели φ_{wt}, θ_{td}:

$$p(s\,|\,d) = \sum_{t\in T} p(t\,|\,d) \prod_{w\in s} p(w\,|\,t)^{n_{sw}} = \sum_{t\in T} \theta_{td} \prod_{w\in s} \varphi_{wt}^{n_{sw}}.$$

Будем считать каждый документ «мешком сегментов». Тогда функция вероятности выборки будет равна произведению функций вероятности сегментов. Поставим задачу максимизации суммы log-правдоподобия и регуляризатора R:

$$\sum_{d\in D} \sum_{s\in S_d} \ln \sum_{t\in T} \theta_{td} \prod_{w\in s} \varphi_{wt}^{n_{sw}} + R(\Phi, \Theta) \to \max_{\Phi,\Theta}, \tag{81}$$

при обычных ограничениях (14).

В частном случае, когда каждый сегмент состоит только из одного слова, данная задача переходит в (15). Заметим также, что задача (81) является частным случаем построения тематической модели гиперграфа (52), в котором вершины являются словами, а рёбра — предложениями.

Теорема 16.1. *Пусть функция $R(\Phi, \Theta)$ непрерывно дифференцируема. Точка (Φ, Θ) локального экстремума задачи* (81), (14) *удовлетворяет системе уравнений со*

вспомогательными переменными $p_{tds} = p(t\,|\,d,s)$, *если из решения исключить нулевые столбцы матриц* Φ, Θ:

$$p_{tds} = \underset{t\in T}{\operatorname{norm}}\big(\theta_{td}\prod_{w\in s}\varphi_{wt}^{n_{sw}}\big);$$

$$\varphi_{wt} = \underset{w\in W}{\operatorname{norm}}\left(n_{wt} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}}\right); \qquad n_{wt} = \sum_{d\in D}\sum_{s\in S_d}[w\in s]\,p_{tds};$$

$$\theta_{td} = \underset{t\in T}{\operatorname{norm}}\left(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right); \qquad n_{td} = \sum_{s\in S_d} p_{tds}.$$

Аналогичным образом задача ставится для модели предложений senLDA [32] и для модели коротких сообщений Twitter-LDA [210]. В обоих случаях регуляризатором являются априорные распределения Дирихле. В модели Twitter-LDA в роли документов выступают авторы, в роли сегментов — сообщения данного автора.

Гиперграфовые модели связного текста. Тематическая модель гиперграфа (52) из главы 10 непосредственно применима для построения тематической модели предложений. Ребром гиперграфа можно описать как предложение, так и любое подмножество термов, не обязательно идущих подряд, связанных друг с другом по смыслу и порождаемых одной общей темой.

Если текст предварительно обработан *синтаксическим парсером*, то в качестве рёбер можно брать ветки или поддеревья синтаксического дерева (*синтагмы*), в частности, *именные группы* — грамматически корректные словосочетания, в которых главным словом является существительное. Синтаксические связи позволяют устанавливать *семантические роли слов* и выделять *факты* в виде троек «объект, субъект, действие», которые можно считать тематически однородными и также описывать рёбрами гиперграфа.

Можно использовать внешние лингвистические ресурсы — *тезаурусы* или *онтологии*, такие как WordNet, Вики-Словарь, РуТез и другие [14]. Они позволяют находить пары терминов, с высокой вероятностью связанные тематически, либо вообще обозначающие в тексте один и тот же объект. Это могут быть пары *синонимов*, пары *гипоним–гипероним* (частное–общее), пары *мероним–холоним* (часть–целое). Термины, связанные тезаурусными отношениями, можно объединять ребром гиперграфа, когда они находятся в одном предложении или в соседних предложениях.

Гиперграфовая модель не накладывает никаких ограничений на многократное вхождение одного и того же слова в разные рёбра гиперграфа. Это даёт возможность учитывать разные типы связей между словами, описывая их различными типами рёбер в гиперграфе. Например, можно в единой гиперграфовой модели учитывать одновременно предложения, факты, синтаксические и тезаурусные связи.

Тематическая сегментация. Теперь рассмотрим более сложный случай, когда текст состоит из предложений, и требуется объединить их в более крупные тематически однородные сегменты, границы которых заранее не определены.

Метод TopicTiling [153] основан на пост-обработке распределений $p(t\,|\,d,w_i)$, $i=1,\ldots,n_d$, получаемых какой-либо тематической моделью, в частности, LDA. Тематика предложения $p(t\,|\,s)$ определяется как средняя тематика $p(t\,|\,d,w)$ всех его

слов[6]. Затем для каждой пары соседних предложений (s_k, s_{k+1}) вычисляется косинусная близость между тематическими векторами $p(t|s_k)$ и $p(t|s_{k+1})$. Чем глубже локальный минимум близости, тем выше уверенность, что между данной парой предложений проходит граница сегментов.

Метод TopicTiling использует набор эвристик для подбора числа предложений слева и справа от локального минимума близости, определения числа сегментов, подбора числа тем и числа итераций, игнорирования стоп-слов, фоновых тем и коротких предложений. Аккуратная настройка параметров этих эвристик позволяет достичь высокого качества сегментации [153].

TopicTiling не является полноценной тематической моделью сегментации текста, поскольку пост-обработка никак не влияет на сами темы. Чтобы найти темы, наиболее выгодные для сегментации, требуется специальный регуляризатор.

Регуляризатор E-шага. До сих пор для формализации дополнительных требований к тематической модели мы конструировали критерии регуляризации как функции от параметров φ_{wt} и θ_{td}. Чтобы учитывать порядок слов внутри документов, гораздо удобнее накладывать ограничения на распределения $p_{tdw} = p(t|d,w)$, которые вычисляются на E-шаге для последовательности слов $(w_1, \ldots, w_{n_d})$ документа d.

С помощью такого регуляризатора можно формализовать требования разреженности тематики предложений или сходства тематики термов внутри каждого предложения, между соседними предложениями, между предложениями одной секции документа, и другие.

Пусть регуляризатор $R(\Pi, \Phi, \Theta)$ является функций не только матриц Φ и Θ, но и трёхмерной матрицы вспомогательных переменных $\Pi = (p_{tdw})_{T\times D\times W}$.

Будем предполагать, что R является достаточно гладкой функцией всех переменных p_{tdw}, φ_{wt} и θ_{td}. Кроме того, сделаем естественное допущение, что если слова w нет в документе d, то функция R не зависит от переменной p_{tdw}.

Согласно уравнению (16), матрица Π является функцией от 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 log-правдоподобия с регуляризацией при ограничениях неотрицательности и нормировки (14):

$$\sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{t\in T} \varphi_{wt}\theta_{td} + R(\Pi(\Phi,\Theta),\Phi,\Theta) \;\to\; \max_{\Phi,\Theta}. \tag{82}$$

Теорема 16.2. *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 (82), (14) удовлетворяет системе уравнений со вспомогательными переменными p_{tdw}*

[6] Точнее, в [153] предлагалось для каждого слова выбирать наиболее вероятную тему, затем усреднять тематику по всем словам предложения. Оба варианта имеют право на существование. Какой из них лучше, в данной работе не исследовалось.

и $\tilde{p}_{tdw}$, если из решения исключить нулевые столбцы матриц Φ, Θ:

$$p_{tdw} = \underset{t\in T}{\operatorname{norm}}\big(\varphi_{wt}\theta_{td}\big); \tag{83}$$

$$\tilde{p}_{tdw} = p_{tdw}\left(1 + \frac{1}{n_{dw}}\left(\frac{\partial R}{\partial p_{tdw}} - \sum_{z\in T} p_{zdw}\frac{\partial R}{\partial p_{zdw}}\right)\right); \tag{84}$$

$$\varphi_{wt} = \underset{w\in W}{\operatorname{norm}}\left(\sum_{d\in D} n_{dw}\tilde{p}_{tdw} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}}\right); \tag{85}$$

$$\theta_{td} = \underset{t\in T}{\operatorname{norm}}\left(\sum_{w\in d} n_{dw}\tilde{p}_{tdw} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right). \tag{86}$$

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

$$\begin{aligned}\varphi_{wt}\frac{\partial p_{zdw}}{\partial \varphi_{wt}} &= \varphi_{wt}\frac{\partial}{\partial \varphi_{wt}}\left(\frac{\varphi_{wz}\theta_{zd}}{\sum_u \varphi_{wu}\theta_{ud}}\right) = \\ &= \varphi_{wt}\frac{[z{=}t]\theta_{td}\sum_u \varphi_{wu}\theta_{ud} - \theta_{td}\varphi_{wz}\theta_{zd}}{(\sum_u \varphi_{wu}\theta_{ud})^2} = p_{tdw}[z{=}t] - p_{tdw}p_{zdw};\end{aligned} \tag{87}$$

$$\begin{aligned}\theta_{td}\frac{\partial p_{zdw}}{\partial \varphi_{td}} &= \theta_{td}\frac{\partial}{\partial \theta_{td}}\left(\frac{\varphi_{wz}\theta_{zd}}{\sum_u \varphi_{wu}\theta_{ud}}\right) = \\ &= \theta_{td}\frac{[z{=}t]\varphi_{wt}\sum_u \varphi_{wu}\theta_{ud} - \varphi_{wt}\varphi_{wz}\theta_{zd}}{(\sum_u \varphi_{wu}\theta_{ud})^2} = p_{tdw}[z{=}t] - p_{tdw}p_{zdw}.\end{aligned} \tag{88}$$

Заметим, что результирующие выражения (87) и (88) совпадают.

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 $w \notin d$:

$$\frac{\partial \tilde{R}}{\partial \varphi_{wt}} = \frac{\partial R}{\partial \varphi_{wt}} + \sum_{z,d,w'} \frac{\partial R}{\partial p_{zdw'}}\frac{\partial p_{zdw'}}{\partial \varphi_{wt}} = \frac{\partial R}{\partial \varphi_{wt}} + \sum_{d\in D}\sum_{z\in T} \frac{\partial R}{\partial p_{zdw}}\frac{\partial p_{zdw}}{\partial \varphi_{wt}}; \tag{89}$$

$$\frac{\partial \tilde{R}}{\partial \theta_{td}} = \frac{\partial R}{\partial \theta_{td}} + \sum_{z,d',w} \frac{\partial R}{\partial p_{zd'w}}\frac{\partial p_{zd'w}}{\partial \theta_{td}} = \frac{\partial R}{\partial \theta_{td}} + \sum_{w\in d}\sum_{z\in T} \frac{\partial R}{\partial p_{zdw}}\frac{\partial p_{zdw}}{\partial \theta_{td}}. \tag{90}$$

Воспользовавшись (87) и (88), получим два тождества:

$$\left.\begin{aligned}\varphi_{wt}\sum_{z\in T}\frac{\partial R}{\partial p_{zdw}}\frac{\partial p_{zdw}}{\partial \varphi_{wt}} \\ \theta_{td}\sum_{z\in T}\frac{\partial R}{\partial p_{zdw}}\frac{\partial p_{zdw}}{\partial \theta_{td}}\end{aligned}\right\} = \sum_{z\in T}\frac{\partial R}{\partial p_{zdw}}p_{tdw}\big([z{=}t] - p_{zdw}\big) = p_{tdw}Q_{tdw},$$

где введена вспомогательная переменная $Q_{tdw} = \dfrac{\partial R}{\partial p_{tdw}} - \sum\limits_{z\in T} p_{zdw}\dfrac{\partial R}{\partial p_{zdw}}$.

Подставим полученное выражение $p_{tdw}Q_{tdw}$ в правую часть (89), домноженную на φ_{wt}, и в правую часть (90), домноженную на θ_{td}. Затем подставим частные произ-

водные регуляризатора $\tilde{R}(\Phi,\Theta)$ в систему уравнений из теоремы 4.1:

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

$$\varphi_{wt} = \underset{w\in W}{\operatorname{norm}}\left(\sum_{d\in D} n_{dw}p_{tdw} + \sum_{d\in D} Q_{tdw}p_{tdw} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}}\right); \tag{91}$$

$$\theta_{td} = \underset{t\in T}{\operatorname{norm}}\left(\sum_{w\in d} n_{dw}p_{tdw} + \sum_{w\in d} Q_{tdw}p_{tdw} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right). \tag{92}$$

Выделение вспомогательной переменной $\tilde{p}_{tdw}$ согласно (84) позволяет переписать уравнения (91)–(92) в требуемом виде (85)–(86).

Теорема доказана.

Таким образом, в EM-алгоритме для каждого документа d сначала вычисляются вспомогательные переменные p_{tdw}, затем они преобразуются в новые переменные $\tilde{p}_{tdw}$, которые подставляются в формулы M-шага (17)–(18) вместо p_{tdw}. Такой способ вычислений будем называть *регуляризацией E-шага* или *пост-обработкой E-шага*.

Заметим, что переменные $\tilde{p}_{tdw}$ могут принимать отрицательные значения, поэтому в общем случае они не образуют вероятностных распределений. Тем не менее, для них выполнено условие нормировки $\sum_t \tilde{p}_{tdw} = 1$.

Разреживание распределений $p(t\,|\,d,w)$. Потребуем, чтобы каждый терм в документе относился к небольшому числу тем. Для этого будем разреживать распределения $p(t\,|\,d,w)$, максимизируя их KL-дивергенции с равномерным распределением:

$$\mathrm{KL}\big(\tfrac{1}{|T|} \,\big\|\, p(t\,|\,d,w)\big) \to \max.$$

Суммируя по всем термам всех документов, получим регуляризатор:

$$R(\Pi) = -\frac{\tau}{|T|}\sum_{d\in D}\sum_{w\in d} n_{dw}\sum_{t\in T}\ln p_{tdw} \to \max.$$

Подставим производную

$$\frac{\partial R(\Pi)}{\partial p_{zdw}} = -\frac{\tau}{|T|}\frac{n_{dw}}{p_{zdw}}$$

в формулу (84):

$$\tilde{p}_{tdw} = p_{tdw} - \tau\big(\tfrac{1}{|T|} - p_{tdw}\big).$$

Таким образом, если для некоторой темы $p_{tdw} < \frac{1}{|T|}$, то на следующей итерации вероятность p_{tdw} для данного терма w станет ещё меньше. Тематика терма будет постепенно концентрироваться в небольшом числе тем.

Ещё одна интерпретация этого регуляризатора следует из возможности записать регуляризацию E-шага эквивалентным образом через формулы M-шага (91)–(92):

$$\varphi_{wt} = \underset{w\in W}{\operatorname{norm}}\left(n_{wt} - \tau n_w\left(\frac{1}{|T|} - \frac{n_{wt}}{n_w}\right)\right);$$

$$\theta_{td} = \underset{t\in T}{\operatorname{norm}}\left(n_{td} - \tau n_d\left(\frac{1}{|T|} - \frac{n_{td}}{n_d}\right)\right).$$

Если нерегуляризованные частотные оценки условных вероятностей $\hat{p}(t\,|\,w) = \frac{n_{wt}}{n_w}$ и $\hat{p}(t\,|\,d) = \frac{n_{td}}{n_d}$ становятся меньше вероятности равномерного распределения $\frac{1}{|T|}$, то происходит разреживание распределений φ_{wt} и θ_{td}; с итерациями их значения уменьшаются и могут обращаться в нуль. Таким образом, происходит согласованное разреживание матриц Φ и Θ, под управлением одного общего коэффициента регуляризации τ.

Разреживающий регуляризатор E-шага для сегментации. Применим регуляризацию E-шага для построения тематической модели сегментированного текста. Сегментами могут быть абзацы, предложения или синтаксически связанные части предложений, найденные с помощью синтаксического анализатора. Обозначим через S_d множество сегментов, на которые разбит документ d, через n_s — длину сегмента s, через n_{sw} — число вхождений терма w в сегмент s.

Определим тематику сегмента $s \in S_d$ как среднюю тематику всех его термов:

$$p_{tds} \equiv p(t\,|\,d,s) = \sum_{w \in s} p(t\,|\,d,w)\,p(w\,|\,s) = \frac{1}{n_s}\sum_{w \in s} n_{sw} p_{tdw}.$$

Чтобы каждый сегмент относился к небольшому числу тем, будем минимизировать кросс-энтропию между распределениями $p(t\,|\,d,s)$ и равномерным распределением, что приведёт нас к разреживающему регуляризатору E-шага:

$$R(\Pi) = -\tau \sum_{d \in D} \sum_{s \in S_d} \sum_{t \in T} \ln \sum_{w \in s} n_{sw} p_{tdw}. \tag{93}$$

Опуская выкладки, приведём результат подстановки (93) в (84):

$$\tilde{p}_{tdw} = p_{tdw}\left(1 - \frac{\tau}{n_{dw}} \sum_{s \in S_d} \frac{n_{sw}}{n_s}\left(\frac{1}{p_{tds}} - \sum_{z \in T} \frac{p_{zdw}}{p_{zds}}\right)\right).$$

Хотя формула выглядит громоздкой, эффект применения регуляризатора понять не трудно. Если вероятность p_{tds} темы в сегменте окажется меньше некоторого порога, то вероятности p_{tdw} будут уменьшаться для всех термов w данного сегмента. В итоге тематика каждого сегмента сконцентрируется в небольшом числе тем.

В результате разреживания тематика соседних сегментов может оказаться близкой, и их можно будет объединить в один тематический сегмент. Назовём тему t с максимальным значением $p(t\,|\,d,s)$ *доминирующей темой* сегмента s документа d. Если тема доминирует в соседних сегментах, то она будет доминирующей и в их объединении. Если объединить последовательные сегменты с одинаковой доминирующей темой в один более крупный сегмент, то данная тема также останется в нём доминирующей. Это простая агломеративная стратегия тематической сегментации. В отличие от TopicTiling, у неё нет эвристических параметров, которые надо настраивать, и она почти не увеличивает время пост-обработки E-шага.

Эвристическая пост-обработка E-шага эквивалентна регуляризации. Согласно формуле (84), произвольному гладкому регуляризатору $R(\Pi, \Phi, \Theta)$ однозначно соответствует преобразование $p_{tdw} \to \tilde{p}_{tdw}$. Верно и обратное: совершив преобразование $p_{tdw} \to \tilde{p}_{tdw}$ с помощью некоторого эвристического алгоритма, можно указать регуляризатор $R(\Pi)$, который приводит к такому же результату.

Теорема 16.3. *Если на k-й итерации EM-алгоритма для каждого (d,w): $n_{dw}>0$ в формулах M-шага вместо вектора $(p^k_{tdw})_{t\in T}$ подставить другой вектор $(\tilde{p}^k_{tdw})_{t\in T}$, удовлетворяющий условию нормировки $\sum_t \tilde{p}^k_{tdw}=1$, то это эквивалентно применению регуляризатора сглаживания–разреживания*

$$R(\Pi)=\sum_{d\in D}\sum_{w\in d} n_{dw}\sum_{t\in T}(\tilde{p}^k_{tdw}-p^k_{tdw})\ln p_{tdw}. \tag{94}$$

Доказательство. Запишем (84) как систему дифференциальных уравнений относительно функции R и выделим в каждом уравнении переменные x_{tdw}:

$$\underbrace{p^k_{tdw}\frac{\partial R}{\partial p_{tdw}}}_{x_{tdw}} - p^k_{tdw}\sum_{z\in T}\underbrace{p^k_{zdw}\frac{\partial R}{\partial p_{zdw}}}_{x_{zdw}} = n_{dw}(\tilde{p}^k_{tdw}-p^k_{tdw}),\quad t\in T.$$

Для каждого (d,w) такого, что $n_{dw}>0$, это система $|T|$ линейных уравнений относительно $|T|$ переменных x_{tdw}, $t\in T$. Сумма всех уравнений равна нулю, следовательно, система является недоопределённой и имеет бесконечное множество решений. Перепишем эту систему в следующем виде:

$$x_{tdw}=n_{dw}(\tilde{p}^k_{tdw}-p^k_{tdw})+p^k_{tdw}\sum_{z\in T}x_{zdw},\quad t\in T.$$

Вектор $x_{tdw}=n_{dw}(\tilde{p}^k_{tdw}-p^k_{tdw})$, $t\in T$ является неподвижной точкой и решением этой системы, поскольку для него $\sum_{z\in T}x_{zdw}=0$. Взяв это решение (или любое другое), получим систему дифференциальных уравнений в частных производных относительно функции $R(\Pi)$:

$$\frac{\partial R}{\partial p_{tdw}}=\frac{x_{tdw}}{p_{tdw}},\quad d\in D,\ w\in d,\ t\in T.$$

Эта система полностью декомпозируется по переменным p_{tdw}. Её решением при заданных d,w,t являются функции вида $R(\Pi)=x_{tdw}\ln p_{tdw}+C$, где C не зависит от переменной p_{tdw}. Следовательно, общее решение $R(\Pi)$ представимо в виде суммы

$$R(\Pi)=\sum_{d\in D}\sum_{w\in d}\sum_{t\in T}x_{tdw}\ln p_{tdw},$$

в которую можно подставить произвольное решение x_{tdw}, $t\in T$ системы линейных уравнений, в том числе и найденную выше неподвижную точку.

Теорема доказана.

На стадии эвристической пост-обработки матрица $\Pi_d=(p_{tdw_i})$ размера $T\times n_d$ рассматривается как последовательность из n_d векторов размерности $|T|$ или пучок из $|T|$ временных рядов длины n_d. На рис. 24 показан в качестве примера возможный вариант пост-обработки. Сначала документ секционируется по предложениям и в каждом предложении определяется доминирующая тема. Затем соседние предложения с одинаковой доминирующей темой объединяются в более крупные сегменты s, для которых определяются распределения $p(t\,|\,d,s)$.

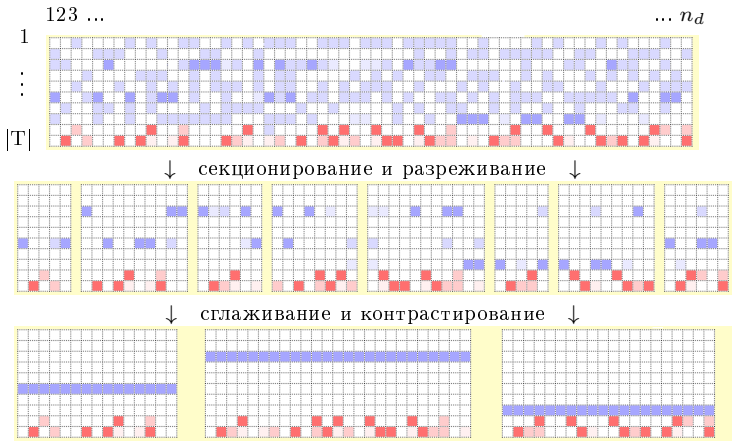


Рис. 24: Эвристическая пост-обработка матрицы $\Pi_d = \big(p(t\,|\,d,w_i)\big)_{T\times n_d}$ на E-шаге для выделения тематически однородных сегментов текстового документа $d = (w_1,\ldots,w_{n_d})$. Предметные темы (синие) подвергаются сглаживанию и разреживанию, затем контрастируются доминирующие темы сегментов. Фоновые темы (две последние, красные) не участвуют в пост-обработке.

Выводы по главе Всего мы рассмотрели пять механизмов тематического моделирования связного текста. Попробуем их ранжировать, начиная с наиболее сильных, учитывающих порядок слов полностью, заканчивая более слабыми, учитывающими его лишь частично или в меньшей степени.

- *Линейная тематизация с локальным E-шагом* из главы 15. Вместо тематики документа $p(t\,|\,d)$ вычисляется тематика локального контекста $p(t\,|\,i)$ в каждой позиции i, затем строятся контекстные тематические векторы $p(t\,|\,d,w_i)$.

- *Пост-обработка E-шага.* Тематические векторы $p(t\,|\,d,w_i)$ сначала вычисляются независимо для каждого слова w_i, затем их последовательность подвергается любому разумному преобразованию, что эквивалентно регуляризации E-шага.

- *Гиперграфовая тематическая модель.* Сочетания термов, предположительно порождаемые одной общей темой (предложения, фразы, факты, и т. д.), объединяются в рёбра гиперграфа. Коллекция остаётся «мешком сочетаний».

- *Тематическая модель сети слов* (или битермов) из главы 14. По каждому слову формируется и добавляется в коллекцию псевдодокумент, составленный путём объединения всех локальных контекстов данного слова как «мешков слов».

- *Тематическая модель n-грамм* из главы 14. Словари n-грамм строятся на этапе предобработки коллекции и используются как модальности. Коллекция остаётся «мешком униграмм и n-грамм». Улучшается интерпретируемость тем.

17 Критерии качества тематических моделей

Критерии качества тематических моделей принято делить на внутренние (intrinsic) и внешние (extrinsic). *Внутренние критерии* характеризуют качество модели по исходной текстовой коллекции. *Внешние критерии* оценивают полезность модели с точки зрения приложения и конечных пользователей. Иногда для этого приходится собирать дополнительные данные, например, оценки асессоров.

На практике к тематическим моделям предъявляются различные наборы требований, а для построения модели применяется многокритериальная оптимизация. Поэтому и качество модели должно оцениваться по многим критериям.

В ARTM избыточная регуляризация может приводить к деградации модели. Образно говоря, регуляризаторы, как лекарства для модели, требуют подбора терапевтической дозы воздействия (коэффициента регуляризации), а в случае передозировки могут приводить в вырождению модели. Для обнаружения нежелательной симптоматики нужны разнообразные критерии качества.

В проекте `BigARTM` поддерживается библиотека стандартных метрик качества и предусмотрены механизмы добавления новых пользовательских метрик.

Внешние критерии весьма разнообразны и зависят от конечной решаемой задачи. Практически в каждой публикации по тематическому моделированию используется какой-либо внешний критерий: качество классификации документов [156], точность и полнота информационного поиска [203, 23, 10, 20], число найденных хорошо интерпретируемых тем [26], качество сегментации текстов [153]. В [50] предлагается методика диагностики тематических моделей, основанная на сопоставлении найденных тем со специальными наборами взаимосвязанных слов, называемыми *концептами*.

Перплексия. Наиболее распространённым внутренним критерием является *перплексия* (perplexity), используемая для оценивания вероятностных моделей языка в компьютерной лингвистике. Это мера несоответствия или «удивлённости» модели $p(w\,|\,d)$ термам w, которые встречаются в документах d. Она определяется через log-правдоподобие (13), либо через log-правдоподобие (39) каждой модальности m:

$$\mathscr{P}_m(D;p) = \exp\left(-\frac{1}{n_m}\sum_{d\in D}\sum_{w\in W^m} n_{dw}\ln p(w\,|\,d)\right), \tag{95}$$

где $n_m = \sum\limits_{d\in D}\sum\limits_{w\in W^m} n_{dw}$ — длина коллекции по m-й модальности. Чем меньше перплексия, тем лучше модель p предсказывает появление термов w в документах d.

Перплексия имеет следующую интерпретацию. Если термы w порождаются из равномерного распределения $p(w) = 1/V$ на словаре мощности V, то перплексия языковой модели $p(w)$ на таком тексте сходится к V с ростом его длины. Чем сильнее распределение $p(w)$ отличается от равномерного, тем меньше перплексия. В случае условных вероятностей $p(w\,|\,d)$ интерпретация немного другая: если каждый документ генерируется из V равновероятных термов (возможно, различных в разных документах), то перплексия сходится к V.

Недостатком перплексии является неочевидность её числовых значений, а также её зависимость от размерных характеристик коллекции — длины документов, мощности словаря, разреженности вероятностного распределения термов. В частности,

с помощью перплексии некорректно сравнивать тематические модели одной и той же коллекции, построенные на разных словарях.

Обозначим через $p_D(w\,|\,d)$ модель, построенную по обучающей коллекции документов D. Перплексия на обучении $\mathscr{P}_m(D; p_D)$ является оптимистично смещённой (заниженной) оценкой качества модели из-за эффекта переобучения. Обобщающую способность тематических моделей принято оценивать *перплексией контрольной выборки* (hold-out perplexity) $\mathscr{P}_m(D'; p_D)$. Обычно коллекцию разделяют на обучающую и контрольную случайным образом в пропорции $9:1$ [41].

В ранних экспериментах было показано, что LDA существенно превосходит PLSA по перплексии, откуда был сделан вывод, что LDA меньше переобучается [41]. Позже было показано, что на больших коллекциях перплексия моделей PLSA и LDA отличается незначительно [115, 200, 109].

На самом деле природа «переобучения» больше связана с особенностями перплексии, чем с качеством самих моделей. Перплексия чувствительна к малым значениям предсказанной вероятности термов, поскольку $p(w\,|\,d)$ стоит под логарифмом. Сглаживание в LDA завышает оценки вероятностей редких термов, поэтому LDA имеет меньшую контрольную перплексию. Моделирование вероятности редких слов важно для статистического машинного перевода и других приложений компьютерной лингвистики, откуда перплексия и пришла в тематическое моделирование. Однако для понимания тематической кластерной структуры текстовой коллекции и выявления тематики отдельных документов редкие термы как раз наименее важны.

В [6, 147] были предложены *робастные тематические модели*, описывающие редкие термы специальным «фоновым» распределением. Контрольная перплексия робастных вариантов PLSA и LDA оказалась существенно меньшей и практически одинаковой.

В [7] было показано, что на достаточно больших коллекциях ($n > 10^6$) обучающая и контрольная перплексия одинаково ранжируют сравниваемые модели, то есть приводят к одинаковым качественным выводам. Таким образом, для сравнения моделей нет особой необходимости вычислять контрольную перплексию.

При вычислении перплексии может возникнуть проблема нулевой вероятности $p(w\,|\,d) = 0$. Терм w в документе d встречается, тем не менее, модель предсказывает для него нулевую вероятность. В частности, при вычислении контрольной перплексии в документе может встретиться новый терм, который ни разу не встретился при обучении. В таких случаях в перплексии возникает бесконечно большое слагаемое ($-\ln 0 = +\infty$). Простейшее решение этой проблемы заключается в том, чтобы проигнорировать все такие термы и посчитать их долю в коллекции как ещё одну меру качества модели. Другое решение — следуя [7], считать все такие термы нетематическими и описывать их вероятность частотной оценкой $p(w\,|\,d) = \frac{n_w}{n}$. Похожий результат получится при использовании тематической модели с необучаемой фоновой темой $p(w\,|\,b) = \frac{n_w}{n}$. Такая модель никогда не будет давать нулевую вероятность терма, если, конечно, вероятность фоновой темы в документе не равна нулю.

Интерпретируемость тематической модели является плохо формализуемым требованием. Содержательно оно означает, что по спискам наиболее частотных слов и документов темы эксперт может понять, о чём эта тема, и дать ей адекватное название [45]. Примеры хорошо интерпретируемых тем показаны на рис. 13 и рис. 21. Свойство интерпретируемости важно в информационно-поисковых системах, исполь-

зующих автоматически найденные темы как инструмент визуализации результатов поиска, например для вывода пояснений или сниппетов, либо как инструмент рубрикации или навигации по текстовой коллекции.

Большинство существующих методов оценивания интерпретируемости основано на привлечении экспертов-асессоров. В [132] экспертам предлагалось непосредственно оценивать полезность тем по трёхбалльной шкале, рассматривая списки слов, ранжированные по убыванию $p(w|t)$. В *методе интрузий* [45] для каждой темы составляется список из 10 верхних слов списка, в который искусственно внедряется одно случайное слово. Тема считается интерпретируемой, если подавляющее большинство экспертов правильно указывают лишнее слово.

В прикладных социологических исследованиях [42, 93, 137] для экспертного оценивания темы используются не только списки верхних слов, но и списки документов, ранжированные по убыванию $p(d|t)$. Эта методика более трудоёмка, поскольку эксперт прочитывает документы. Но она более надёжна в тех случаях, когда прикладной целью тематического моделирования является поиск тем определённой направленности (например, обсуждений межэтнических отношений в социальных сетях), затем качественное понимание семантики каждой темы (в частности, какие этничности и какие проблемы затрагивает каждая тема), и наконец количественное оценивание объёма данной темы (где, когда и как часто возникает данный дискурс) [25, 26].

Когерентность. Экспертные подходы необходимы на стадии исследований, но они затрудняют автоматическое построение хороших тематических моделей. В серии работ [132, 133, 134, 125] было показано, что среди критериев качества, вычисляемых по коллекции автоматически, согласованность или *когерентность* (coherence) лучше всего коррелирует с экспертными оценками интерпретируемости.

Тема называется *когерентной* (согласованной), если термы, наиболее частые в данной теме, неслучайно часто совместно встречаются рядом в документах коллекции [133, 134]. Численной мерой когерентности темы t является *поточечная взаимная информация* (71), вычисляемая по k наиболее вероятным словам темы:

$$\mathscr{C}_t = \frac{2}{k(k-1)} \sum_{i=1}^{k-1} \sum_{j=i}^{k} \mathrm{PMI}(w_i, w_j), \tag{96}$$

где w_i — i-й терм в порядке убывания φ_{wt}, число k обычно полагается равным 10.

Когерентность модели определяется как средняя когерентность $\mathscr{C}_t$ по всем темам. Когерентность может оцениваться по сторонней коллекции (например, по Википедии) [130], либо по той же коллекции, по которой строится модель [125].

Разреженность и семантические ядра тем. Разреженность модели измеряется долей нулевых элементов в матрицах Φ и Θ. В моделях, разделяющих множество тем T на предметные S и фоновые B, разреженность оценивается только по столбцам Φ и строкам Θ, соответствующим предметным темам.

Недостаток такого определения разреженности в его неустойчивости. На практике матрицы Φ и Θ могут содержать большую долю значений, близких к нулю. Их обнуление практически не повлияет на модель, но резко повысит разреженность.

В [93] предлагается считать *существенными* лишь те условные вероятности, значения которых выше, чем у равномерного распределения:

$$W_t = \{w \in W \mid \varphi_{wt} > \tfrac{1}{|W|}\};$$
$$T_d = \{t \in T \mid \theta_{td} > \tfrac{1}{|T|}\}.$$

В экспериментах на 300 тысячах постов социальной сети при 120 темах разреженность Φ превысила 96%, разреженность Θ — 88%. При этом число слов, вошедших хотя бы в одно из множеств W_t, оказалось равным 8 тысячам при словаре 154 тысячи слов. Такой подход позволяет сократить словарь до минимального набора тематически релевантных слов. Если требуется управлять разреженностью модели, то придётся вводить условия $\varphi_{wt} > \varphi_0$, $\theta_{td} > \theta_0$ и подбирать пороги φ_0, θ_0.

В [180] предлагается определять *семантическое ядро темы* как множество слов, которые с большой вероятностью употребляются в теме t и редко употребляются в других темах: $W_t = \{w \in W \mid p(t\,|\,w) > 0.25\}$, где $p(t\,|\,w) = \varphi_{wt}\frac{n_t}{n_w}$. Порог 0.25 был подобран эмпирическим путём при числе тем $|T| = 100$. Для других данных и другого числа тем, его, возможно, придётся подбирать заново.

Независимо от того, каким образом определяется семантическое ядро, введём несколько показателей, характеризующих разреженность матрицы Φ:

$$\ker_t = |W_t| \text{ — } \textit{размер ядра} \text{ (оптимум немного выше } \tfrac{|W|}{|T|});$$
$$\operatorname{pur}_t = \sum_{w \in W_t} p(w\,|\,t) \text{ — } \textit{чистота темы} \text{ (чем выше, тем лучше);}$$
$$\operatorname{con}_t = \tfrac{1}{|W_t|}\sum_{w \in W_t} p(t\,|\,w) \text{ — } \textit{контрастность темы} \text{ (чем выше, тем лучше);}$$
$$\operatorname{logLift}_t = \tfrac{1}{|W_t|}\sum_{w \in W_t} \log \tfrac{p(w\,|\,t)}{p(w)} \text{ — } \textit{подъём темы} \text{ (чем выше, тем лучше).}$$

Для модели в целом эти показатели определяются путём усреднения по всем предметным темам $t \in S$. Косвенно они являются также и мерой интерпретируемости модели, поскольку интерпретируемые темы должны обладать не слишком маленьким, но и не слишком большим семантическим ядром.

Логарифмический критерий чистоты (подъём) темы logLift показывает, на сколько порядков частота термов из семантического ядра темы превышает их обычную частоту в коллекции $p(w)$. Он был предложен в [170] для сортировки термов при визуализации темы в пользовательском интерфейсе. В [65] было предложено усреднять logLift по $k = 30$ ключевым словам в каждой теме. Было показано, что средний logLift связан с долей информативных слов в темах, и что он значимо коррелирует с экспертными оценками качества тем.

Доля фоновой лексики. Пусть $B \subset T$ — подмножество фоновых тем, в которых собрана общеупотребительная лексика. Определим *долю фоновой лексики* для всей коллекции и для отдельного документа d:

$$\mathscr{B} = \frac{1}{n}\sum_{d \in D}\sum_{w \in d}\sum_{t \in B} n_{dw} p(t\,|\,d, w);$$
$$\mathscr{B}_d = \frac{1}{n_d}\sum_{w \in d}\sum_{t \in B} n_{dw} p(t\,|\,d, w).$$

Доля фоновой лексики принимает значения от 0 до 1. Если она близка к 0, то модель не способна выделять слова общей лексики, если же она близка к 1, то это свидетельствует о вырождении модели. В специализированных текстах узких предметных областей доля фоновой лексики может составлять от 30% до 90%.

Возможна ситуация, когда среднее по коллекции значение $\mathscr{B}$ укладывается в эти ориентировочные нормы, однако $\mathscr{B}_d$ выходит за их пределы для значительной доли документов. Это может говорить как о наличии аномальных или «мусорных» документов в коллекции, так и о недостатках моделирования. Анализ аномальных документов может привести к пониманию этих недостатков с последующей модификацией модели, оптимизационных критериев или регуляризаторов.

Такие критерии, как разреженность, размер ядра или доля фоновой лексики могут использоваться для контроля избыточного разреживания модели.

Различность тем. Введём функцию расстояния между темами $\rho(t,s)$ как распределениями $p(w\,|\,t)$ и $p(w\,|\,s)$, то есть столбцами матрицы Ф. Определим для каждой темы t расстояние до ближайшей к ней темы s:

$$\mathscr{R}_t = \min_{s \in T \setminus t} \rho(t,s).$$

Если расстояние $\mathscr{R}_t$ до ближайшей темы s близко к нулю, то темы t и s можно считать дубликатами. Возможно, эти темы следовало бы объединить.

Если расстояния $\mathscr{R}_t$ велики для всех тем, то это означает, что все темы попарно существенно различны, то есть модель *хорошо декоррелирована*.

Расстояния между темами как столбцами матрицы Ф можно определять по-разному. Обычно используется одно из следующих:

$$\rho(t,s) = 1 - \frac{\sum_w \varphi_{ws}\varphi_{wt}}{\big(\sum_w \varphi_{ws}^2\big)^{1/2}\big(\sum_w \varphi_{wt}^2\big)^{1/2}} \quad \text{— косинусное расстояние;} \tag{97}$$

$$\rho(t,s) = \Big(\tfrac{1}{2}\sum_w \big(\sqrt{\varphi_{ws}} - \sqrt{\varphi_{wt}}\big)^2\Big)^{1/2} \quad \text{— расстояние Хеллингера;} \tag{98}$$

$$\rho(t,s) = 1 - |W_t \cap W_s| : |W_t \cup W_s| \quad \text{— расстояние Жаккара;} \tag{99}$$

где W_t и W_s — семантические ядра тем.

Выводы по главе

- Задача тематического моделирования некорректно поставлена. Для регуляризации решения в ARTM используется многокритериальная оптимизация. Соответственно, и качество модели должно измеряться многими критериями.
- В прикладных исследованиях обычно ограничиваются измерением перплексии и когерентности. К сожалению, эта плохая практика весьма распространена.
- В данной главе описаны критерии для анализа различных аспектов качества как отдельных тем, так и модели в целом. Мы ограничились лишь внутренними критериями; внешних критериев гораздо больше, но они специфичны для прикладных задач.
- В главе 18 мы рассмотрим семейство критериев для тестирования гипотезы условной независимости — одного из ключевых допущений вероятностного тематического моделирования.

18 Критерии условной независимости

Гипотеза условной независимости является базовым предположением вероятностного тематического моделирования. При построении вероятностных моделей по эмпирическим данным базовые допущения принято проверять после того, как модель построена. Например, анализ регрессионных остатков содержит в своём арсенале десятки тестов для проверки статистических гипотез о свойствах остатков — независимости, некоррелированности, нормальности, нулевого математического ожидания, постоянства дисперсии, и т. д.

Статистический анализ адекватности вероятностных тематических моделей не настолько хорошо разработан, хотя известны отдельные работы в этом направлении [187, 121, 65].

В данной главе мы введём семейство средневзвешенных статистик, позволяющих измерять семантическую однородность темы, согласованность документов и термов с темой. Неожиданно обнаружится, что перплексия тоже принадлежит этому семейству статистик и допускает интересные обобщения.

Гипотеза условной независимости допускает три эквивалентных представления:

$$p(w,d\,|\,t) = p(w\,|\,t)\,p(d\,|\,t); \tag{100}$$

$$p(w\,|\,d,t) = p(w\,|\,t); \tag{101}$$

$$p(d\,|\,w,t) = p(d\,|\,t). \tag{102}$$

На практике все эти распределения неизвестны. После построения модели (и на каждом шаге EM-алгоритма) доступны лишь оценки этих распределений. В каждом из трёх равенств распределение в правой части оценивается по большему объёму данных, чем распределение в левой части. Поэтому в качестве нулевой гипотезы будем брать предположение, что эмпирическое распределение в левой части порождается (согласуется с) вероятностной моделью из правой части.

Несмотря на формальную эквивалентность, три представления (100), (101), (102) приводят к различным конструкциям статистических тестов с разными возможностями применения. Рассмотрим их подробнее.

1. Равенство (100) трансформируется в нулевую гипотезу о том, что для заданной темы t совместное распределение термов в документах порождается факторизованным распределением:

$$H_0: \quad \hat{p}(w,d\,|\,t) \sim p(w\,|\,t)\,p(d\,|\,t). \tag{103}$$

Для проверки данной гипотезы можно использовать статистику взаимной информации, предложенную в [121]:

$$S_t = \mathrm{KL}\big(\hat{p}(w,d\,|\,t)\;\big\|\;p(w\,|\,t)p(d\,|\,t)\big) = \sum_{d\in D}\sum_{w\in d}\hat{p}(w,d\,|\,t)\ln\frac{\hat{p}(w,d\,|\,t)}{p(w\,|\,t)p(d\,|\,t)}.$$

Для получения удобной вычислительной формулы воспользуемся определением условной вероятности: $\hat{p}(w,d\,|\,t) = p(t\,|\,d,w)\,\hat{p}(w\,|\,d)\frac{p(d)}{p(t)}$, $p(d\,|\,t) = p(t\,|\,d)\frac{p(d)}{p(t)}$, формулой

E-шага (16) и частотной оценкой условной вероятности $\hat{p}(w,d\,|\,t) = \frac{n_{tdw}}{n_t}$. Подставим полученные выражения в S_t:

$$\frac{\hat{p}(w,d\,|\,t)}{p(w\,|\,t)\,p(d\,|\,t)} = \frac{p(t\,|\,d,w)\,\hat{p}(w\,|\,d)}{p(w\,|\,t)\,p(t\,|\,d)} = \frac{p_{tdw}}{\varphi_{wt}\theta_{td}}\hat{p}(w\,|\,d) = \frac{\hat{p}(w\,|\,d)}{p(w\,|\,d)};$$

$$S_t = \sum_{d\in D}\sum_{w\in d}\frac{n_{tdw}}{n_t}\ln\frac{\hat{p}(w\,|\,d)}{p(w\,|\,d)} = \operatorname*{avg}_{d,w}\left(n_{tdw}, \ln\frac{\hat{p}(w\,|\,d)}{p(w\,|\,d)}\right), \tag{104}$$

где $\operatorname*{avg}_{i\in I}(\gamma_i, x_i) = \frac{\sum_{i\in I}\gamma_i x_i}{\sum_{i\in I}\gamma_i}$ — средневзвешенное значений x_i с весами γ_i, $i \in I$.

Статистика S_t равна средневзвешенному значению логарифмической функции потерь $\ell(w,d) = \ln\frac{\hat{p}(w\,|\,d)}{p(w\,|\,d)}$ по всем термам w всех документов d, взятым с весами n_{tdw}.

Функция потерь $\ell(w,d)$ положительна, когда $p(w\,|\,d) < \hat{p}(w\,|\,d)$. Потеря тем выше, чем хуже тематическая модель предсказывает появление терма в документе $p(w\,|\,d)$ по сравнению с тривиальной частотной оценкой $\hat{p}(w\,|\,d) = \frac{n_{dw}}{n_d}$.

Статистика S_t является мерой семантической неоднородности темы t. Чем больше S_t, тем хуже тема. Чем больше в модели семантически неоднородных тем, тем хуже модель в целом. Возможно, такую модель следует перестроить, увеличив число тем $|T|$, либо расщепить темы с наибольшими S_t на подтемы в иерархической модели.

2. Равенство (101) трансформируется в нулевую гипотезу о том, что эмпирическое распределение термов темы t в документе d порождается общим для всех документов распределением термов:

$$H_0: \quad \hat{p}(w\,|\,d,t) \sim p(w\,|\,t). \tag{105}$$

Для проверки данной гипотезы относительно фиксированного документа d введём статистику на основе KL-дивергенции и, опуская аналогичные выкладки, запишем её через средневзвешенное логарифмической функции потерь:

$$S_{td} = \mathrm{KL}\big(\hat{p}(w\,|\,d,t)\;\big\|\;p(w\,|\,t)\big) = \operatorname*{avg}_{w\in d}\left(n_{tdw}, \ln\frac{\hat{p}(w\,|\,d)}{p(w\,|\,d)}\right). \tag{106}$$

Статистика S_{td} является мерой несогласованности документа d с темой t.

Чем больше значение S_{td}, тем хуже тема определена в документе. Аномально высокие значения S_{td} могут говорить о недостаточном числе итераций для данного документа или о наличии в нём новых неизвестных тем. Аномально низкие значения S_{td} могут служить критерием отбора документов, наиболее релевантных данной теме при суммаризации или визуализации темы.

3. Равенство (102) трансформируется в нулевую гипотезу о том, что эмпирическое распределение терма w по документам в теме t порождается общим для всех термов распределением $p(d\,|\,t)$:

$$H_0: \quad \hat{p}(d\,|\,w,t) \sim p(d\,|\,t). \tag{107}$$

Для проверки данной гипотезы относительно терма w введём статистику на основе KL-дивергенции. Снова опуская выкладки, запишем её через средневзвешенное логарифмической функции потерь:

$$S_{wt} = \mathrm{KL}\big(\hat{p}(d\,|\,w,t)\;\big\|\;p(d\,|\,t)\big) = \operatorname*{avg}_{d\in D}\left(n_{tdw}, \ln\frac{\hat{p}(w\,|\,d)}{p(w\,|\,d)}\right). \tag{108}$$

Статистика S_{wt} является мерой несогласованности терма w с темой t.

Аномально высокие значения S_{wt} могут говорить о том, что терм относится к общеупотребительной лексике. Для выделения таких термов в фоновые темы можно усилить регуляризаторы декоррелирования и сглаживания [181].

Аномально низкие значения S_{wt} говорят о том, что терм w входит в семантическое ядро темы. Темы, содержащие большое число таких термов, могут быть образованы шаблонными фразами, часто повторяющимися в текстах коллекции [121].

Для проверки статистической гипотезы об условной независимости значение статистики S_* преобразуется в достигаемый уровень значимости p-value $= F(S_*)$, где F — функция распределения статистики S_*, полученная в условиях истинности нулевой гипотезы. Если достигаемый уровень значимости близок к единице, то делается вывод, что данные противоречат нулевой гипотезе. Функция распределения $F(S_*)$ строится по синтетической коллекции, генерируемой путём сэмплирования термов $w \sim p(w\,|\,d)$ тематической моделью [19].

Обобщённые средневзвешенные статистики. Обобщим введённые выше статистики S_t, S_{td}, S_{wt} на случай произвольной функции потерь $\ell(d,w)$:

$$S_t = \operatorname*{avg}_{d,w}\big(n_{tdw}, \ell(d,w)\big) \text{ — неоднородность темы } t \text{ в коллекции;}$$
$$S_{td} = \operatorname*{avg}_{w\in d}\big(n_{tdw}, \ell(d,w)\big) \text{ — несогласованность документа } d \text{ с темой } t;$$
$$S_{wt} = \operatorname*{avg}_{d\in D}\big(n_{tdw}, \ell(d,w)\big) \text{ — несогласованность терма } w \text{ с темой } t.$$

При логарифмической функции потерь $\ell(d,w) = \ln \frac{\hat{p}(w\,|\,d)}{p(w\,|\,d)}$ обобщённые статистики S_t, S_{td}, S_{wt} переходят в (104), (106), (108) соответственно.

В общем случае потребуем, чтобы функция потерь не зависела от темы и чтобы значение $\ell(d,w)$ было тем выше, чем хуже модель предсказывает появление терма w в документе d.

Далее рассмотрим несколько частных случаев этой общей конструкции.

Меры несогласованности, толерантные к повторяемости слов. Гипотеза условной независимости является избыточно сильным допущением. Некоторые языковые явления могут формально нарушать её, даже при условии, что темы остаются семантически однородными.

Например, явление *повторяемости слов* (word burstiness) заключается в том, что слово, встретившись в тексте один раз, имеет большую вероятность встретится ещё [60, 101]. В научно-технических текстах это связано с требованиями единства терминологии. Кроме того, автор может иметь свои лексические предпочтения. Тема может содержать синонимы, из которых каждый автор употребляет лишь некоторые. Тема может иметь несколько аспектов, имеющих характерные лексические различия, однако автор может затрагивать лишь часть аспектов. В результате документ может содержать намного меньше слов темы, при этом некоторые из них будут встречаться намного чаще, чем можно было бы ожидать при строгом выполнении гипотезы условной независимости.

Чтобы сделать статистики S_t, S_{td}, S_{wt} толерантными к повторяемости слов, заменим частоты слов в документах n_{dw} бинарными индикаторами $b_{dw} = [n_{dw} \geqslant 1]$:

$$S_t = \operatorname*{avg}_{d,w}\big(b_{dw}p_{tdw}, \ell(d,w)\big);$$
$$S_{td} = \operatorname*{avg}_{w\in d}\big(b_{dw}p_{tdw}, \ell(d,w)\big);$$
$$S_{wt} = \operatorname*{avg}_{d\in D}\big(b_{dw}p_{tdw}, \ell(d,w)\big).$$

Чтобы функцию потерь также сделать толерантной к повторяемости, будем сравнивать модельную вероятность $p(w\,|\,d)$ с частотной оценкой $\hat{p}(w\,|\,d) = \frac{b_{dw}}{n_d}$, для общности умноженной на параметр α (уменьшение α делает статистики ещё более толерантными к нарушениям нулевой гипотезы):

$$\ell(d,w) = \big[p(w\,|\,d) < \alpha\tfrac{b_{dw}}{n_d}\big].$$

Теперь статистики S_t, S_{td}, S_{wt} принимают значения из отрезка $[0,1]$ и выражают долю термов темы t, для которых модель предсказывает слишком малую вероятность. Благодаря столь универсальной интерпретации статистики с бинарной функцией потерь можно использовать для сравнения тем в моделях с различным числом тем, с различными словарями и даже построенных по различным коллекциям.

Перплексия темы. Заметим, что логарифм перплексии тематической модели можно записать через средневзвешенную функцию потерь $\ell(d,w) = \ln\frac{1}{p(w\,|\,d)}$, причём как для всей коллекции, так и для отдельного документа:

$$\ln\mathscr{P} = \operatorname*{avg}_{d,w,t}\big(n_{tdw}, \ln\tfrac{1}{p(w\,|\,d)}\big) = \operatorname*{avg}_{d,w}\big(n_{dw}, \ln\tfrac{1}{p(w\,|\,d)}\big);$$
$$\ln\mathscr{P}_d = \operatorname*{avg}_{w,t}\big(n_{tdw}, \ln\tfrac{1}{p(w\,|\,d)}\big) = \operatorname*{avg}_{w\in d}\big(n_{dw}, \ln\tfrac{1}{p(w\,|\,d)}\big).$$

По аналогии определим через обобщённые средневзвешенные логарифм *перплексии темы* t как по всей коллекции, так и по отдельному документу:

$$\ln\mathscr{P}_t \;= S_t \;= \operatorname*{avg}_{d,w}\big(n_{tdw}, \ln\tfrac{1}{p(w\,|\,d)}\big);$$
$$\ln\mathscr{P}_{td} = S_{td} = \operatorname*{avg}_{w\in d}\big(n_{tdw}, \ln\tfrac{1}{p(w\,|\,d)}\big).$$

Значение перплексии $\mathscr{P}_{td}$ не определено, если $n_{tdw} = 0$ для всех $w \in d$, то есть если тема полностью отсутствует в документе.

Дивергенция Кресси–Рида. При функции потерь

$$\ell(d,w) = \frac{1}{\lambda(\lambda+1)}\left(\left(\frac{\hat{p}(w\,|\,d)}{p(w\,|\,d)}\right)^{\lambda} - 1\right)$$

обобщённые средневзвешенные статистики переходят в дивергенцию Кресси–Рида. Это параметрическое семейство статистик используется в качестве обобщения статистики хи-квадрат Пирсона для проверки гипотезы о согласии эмпирического распределения с заданным дискретным распределением [53].

При конкретных значениях параметра λ дивергенция Кресси–Рида переходит (с точностью до множителя) в статистику хи-квадрат Пирсона ($\lambda = 1$), дивергенцию Кульбака–Лейблера ($\lambda \to 0$), статистику Фримана–Тьюки или расстояние Хеллингера ($\lambda = -\frac{1}{2}$), модифицированную статистику логарифма отношения правдоподобий ($\lambda \to -1$), модифицированную статистику хи-квадрат Неймана или взвешенное евклидово расстояние ($\lambda = -2$). Все эти статистики являются несимметричными (за исключением расстояния Хеллингера) функциями пары дискретных распределений.

Преимущество такой параметризации в том, что свободой выбора параметра λ можно распорядиться для оптимизации какого-либо внешнего критерия качества.

Выводы по главе

- Систематический подход к оцениванию адекватности построенной тематической модели заключается в проверке основного вероятностного допущения — гипотезы условной независимости.
- Для построения соответствующего статистического теста вводятся обобщённые средневзвешенные статистики S_t, S_{td}, S_{wt}.
- Они позволяют оценить, насколько плоха тема t в целом для модели, в конкретном документе d, для конкретного терма w.
- Ранжирование тем, документов, термов по значениям этих статистик позволяет выявлять как наилучшие, так и наихудшие темы, документы, термы. Результатом может быть лучшее понимание модели или принятие решений о модификации модели.
- Конструкция этих статистик единообразна. Меняя в ней веса и функцию потерь, можно делать их толерантными к определённым отклонениям модели от гипотезы условной независимости.

19 Особенности реализации EM-алгоритма

EM-алгоритм в тематическом моделировании — это метод простых итераций для решения системы уравнений вида (16)–(18) в случае обычной двухматричной модели или (41)–(43) для мультимодальной модели или (53)–(55) для гиперграфовой модели. Реализации итерационного процесса могут отличаться порядком вычислений по формулам E-шага и M-шага. В рациональном варианте (алгоритм 2, стр. 15) каждая итерация выполняется за один проход по всем термам всех документов, в результате которого формируются счётчики n_{wt}, n_{td} и обновляются параметры модели φ_{wt}, θ_{td}. В данной главе обсуждаются приёмы улучшения сходимости и особенности организации этого итерационного процесса для больших текстовых коллекций.

Пакетный алгоритм позволяет обрабатывать коллекции документов, не помещающиеся в оперативную память. Коллекция D разбивается на пакеты D_b, $b = 1, \ldots, B$, каждый из которых хранится в отдельном файле. Пакеты обрабатываются по очереди. Каждый пакет загружается в память, обновляет матрицу Φ и выгружается, освобождая память для обработки следующего пакета.

Функция ProcessBatches обрабатывает за один раз множество пакетов $\{D_b\}$, см. алгоритм 6. Для каждого документа d каждого из пакетов D_b производятся итерации вектора θ_d со встроенным E-шагом при фиксированной матрице Φ. На последней итерации документа обновляются счётчики $\tilde{n}_{wt}$ текущего пакета.

Оффлайновый и онлайновый алгоритм — это две разные стратегии агрегирования счётчиков, полученных от разных пакетов, в итоговых счётчиках n_{wt}.

Оффлайновый алгоритм FitOffline совершает много проходов по коллекции. На каждом проходе счётчики n_{wt} формируются при фиксированной матрице Φ и суммируются по всем документам. Обновление Φ с учётом всех регуляризаторов производится в конце каждого прохода коллекции. Оффлайновый режим ориентирован на обработку относительно небольших коллекций.

Онлайновый алгоритм FitOnline был предложен для модели LDA в [76], позже для модели PLSA в [33]. Он реализован в библиотеках машинного обучения `Vowpal Wabbit`, `Gensim`, `BigARTM` и других, и считается наиболее эффективным методом обучения тематических моделей. Его основная идея заключается в специальной организации последовательности вычислений по формулам E-шага и M-шага. Она не затрагивает механизмы регуляризации и одинаково применима к PLSA, LDA и ARTM. Онлайновый алгоритм делает один проход по коллекции, обновляя матрицу Φ после каждых h пакетов: на шаге 21 счётчики $\tilde{n}_{wt}$, накопленные по h последним пакетам, суммируются со счётчиками n_{wt}, накопленными по всем предыдущим пакетам, с весами $k_{\texttt{decay}}$ и $k_{\texttt{apply}}$. На шаге 22 матрица Φ пересчитывается по обновлённым счётчикам с учётом регуляризаторов. Онлайновый алгоритм ориентирован на обработку больших коллекций или потоков данных.

Весовые коэффициенты $k_{\texttt{decay}}$ и $k_{\texttt{apply}}$ позволяют управлять темпом забывания предыдущих пакетов. Переменная n_{wt} накапливает сумму значений $\tilde{n}_{wt}$ при $k_{\texttt{decay}} = k_{\texttt{apply}} = 1$, среднее арифметическое при $k_{\texttt{decay}} = 1 - \frac{1}{i}$, $k_{\texttt{apply}} = \frac{1}{i}$, экспоненциальное скользящее среднее при $k_{\texttt{decay}} + k_{\texttt{apply}} = 1$. Для алгоритма Online LDA в [76]

Алгоритм 6. Оффлайновый и онлайновый EM-алгоритм для ARTM.

1 **функция** $(\tilde{n}_{wt}) :=$ ProcessBatches (множество пакетов $\{D_b\}$, матрица Φ);
2 $\tilde{n}_{wt} := 0$ для всех $w \in W,\ t \in T$;
3 **для всех** пакетов D_b, **всех** документов $d \in D_b$
4 инициализировать $\theta_{td} := \frac{1}{|T|}$ для всех $t \in T$;
5 **повторять**
6 $n_{tdw} := n_{dw} \operatorname*{norm}_{t \in T}\big(\varphi_{wt}\theta_{td}\big)$ для всех $w \in d,\ t \in T$;
7 $\theta_{td} := \operatorname*{norm}_{t \in T}\big(\sum_{w \in d} n_{tdw} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\big)$ для всех $t \in T$;
8 **пока** θ_d не сойдётся;
9 $\tilde{n}_{wt} := \tilde{n}_{wt} + n_{tdw}$ для всех $w \in d,\ t \in T$;

10 **функция** FitOffline (коллекция $D = \{D_b \colon b \in B\}$);
11 инициализировать φ_{wt} для всех $w \in W,\ t \in T$;
12 **повторять**
13 $(n_{wt}) := \sum_{b=1}^{B}$ ProcessBatches(D_b, Φ);
14 $\varphi_{wt} := \operatorname*{norm}_{w \in W}\big(n_{wt} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}}\big)$ для всех $w \in W,\ t \in T$;
15 **пока** Φ не сойдётся;

16 **функция** FitOnline (коллекция $D = \{D_b \colon b \in B\}$, параметры $k_{\texttt{decay}}, k_{\texttt{apply}}, h$);
17 инициализировать φ_{wt} для всех $w \in W,\ t \in T$;
18 $n_{wt} := 0$ для всех $w \in W,\ t \in T$;
19 **для** $i := 1, \ldots, B/h$
20 $(\tilde{n}_{wt}) :=$ ProcessBatches$(\{D_{hi-h+1}, \ldots, D_{hi}\}, \Phi)$;
21 $n_{wt} := k_{\texttt{decay}} n_{wt} + k_{\texttt{apply}} \tilde{n}_{wt}$ для всех $w \in W,\ t \in T$;
22 $\varphi_{wt} := \operatorname*{norm}_{w \in W}\big(n_{wt} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}}\big)$ для всех $w \in W,\ t \in T$;

предлагалось брать $k_{\texttt{decay}} = 1 - \rho_i$, $k_{\texttt{apply}} = \rho_i$, где $\rho_i = (\tau_0 + i)^{-\kappa}$, значение τ_0 задаются в диапазоне от 64 до 1024, значения κ — от 0.5 до 0.7.

На практике вместо контроля условий сходимости на шаге 8 и шаге 15 часто задают фиксированное число итераций по коллекции и по каждому документу.

В онлайновом алгоритме разбиение коллекции на пакеты и порядок обработки пакетов могут влиять на результат, в отличие от оффлайнового алгоритма. Чтобы уменьшить это влияние, коллекцию разбивают на пакеты случайным образом.

Параллельный алгоритм. Обработка пакетов может выполняться параллельно в несколько потоков как в онлайновом алгоритме, как и в оффлайновом.

В *синхронном алгоритме* обработка следующей порции пакетов не начинается, пока не завершено обновление матрицы Φ на шагах 21–22. Эти задержки приводят к неэффективной загрузке вычислительных ресурсов.

Проблема решается в *асинхронном онлайновом алгоритме* с помощью обновлений с запаздыванием [68]. Пока один процесс занят формированием нового прибли-

число проц.	$\|T\|$	Gensim	Vowpal Wabbit	BigARTM (синхрон)	BigARTM (асинхрон)
1	50	142m (4945)	50m (5413)	42m (5117)	25m (5131)
1	100	287m (3969)	91m (4592)	52m (4093)	32m (4133)
1	200	637m (3241)	154m (3960)	83m (3347)	**53m** (3362)
2	50	89m (5056)		22m (5092)	13m (5160)
2	100	143m (4012)		29m (4107)	19m (4144)
2	200	325m (3297)		47m (3347)	**28m** (3380)
4	50	88m (5311)		12m (5216)	7m (5353)
4	100	104m (4338)		16m (4233)	10m (4357)
4	200	315m (3583)		26m (3520)	**16m** (3634)
8	50	88m (6344)		8m (5648)	5m (6220)
8	100	107m (5380)		10m (4660)	6m (5119)
8	200	288m (4263)		15m (3929)	**10m** (4309)

Рис. 25: Сравнение времени построения тематической модели (в минутах) и перплексии в зависимости от числа процессоров и числа тем для библиотек Gensim, Vowpal Wabbit, `BigARTM` в экспериментах на коллекции из 3.7 миллионов статей Википедии со словарём 100 тысяч слов [92]. В библиотеке Vowpal Wabbit распараллеливание не поддерживалось на момент проведения экспериментов.

жения матрицы Φ, остальные процессы продолжают использовать её предыдущее приближение для обработки пакетов и обновления счётчиков $\tilde{n}_{wt}$.

На рис. 25 приведены результаты сравнения синхронного и асинхронного параллельного онлайнового алгоритма, реализованного в библиотеке `BigARTM`, с популярными реализациями тематической модели LDA [92]. Реализация `BigARTM` заметно эффективнее использует распараллеливание. При этом асинхронный алгоритм примерно в полтора раза быстрее синхронного. На восьми процессорах асинхронный `BigARTM` почти в 30 раз быстрее Gensim.

В обзорной статье [3] описаны 11 технических приёмов для повышения эффективности параллельных алгоритмов тематического моделирования и 14 библиотек, в которых эти приёмы реализованы в различных сочетаниях.

Произвольные функции потерь и E-шаг без нормировки. Изменим критерий оптимизационной задачи, заменив в критерии правдоподобия логарифм на произвольную гладкую неубывающую функцию $\ell(p)$. Теперь для оценивания параметров тематической модели Φ и Θ по коллекции документов D будем максимизировать сумму потерь $\ell\big(p(w\,|\,d)\big)$ по всем терминам во всех документах:

$$\sum_{d\in D}\sum_{w\in d} n_{dw}\ell\Big(\sum_{t\in T}\varphi_{wt}\theta_{td}\Big) + R(\Phi,\Theta) \;\to\; \max_{\Phi,\Theta}. \tag{109}$$

Теорема 19.1. *Решение Φ,Θ задачи (109) при ограничениях неотрицательности и нормировки удовлетворяет системе уравнений со вспомогательными переменными $p_{tdw} = p(t\,|\,d,w)$, если из решения исключить нулевые столбцы матриц Φ, Θ:*

$$p_{tdw} = \varphi_{wt}\theta_{td}\ell'\big(p(w\,|\,d)\big); \tag{110}$$

$$\varphi_{wt} = \underset{w\in W}{\operatorname{norm}}\left(\sum_{d\in D} n_{dw}p_{tdw} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}}\right); \tag{111}$$

$$\theta_{td} = \underset{t\in T}{\operatorname{norm}}\left(\sum_{w\in d} n_{dw}p_{tdw} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right). \tag{112}$$

Система уравнений отличается от классической только формулой E-шага.

Доказательство следует из леммы 3.2 о максимизации на единичных симплексах.

Только при $\ell(p) = \ln p$ на E-шаге возникает формула Байеса. При этом задача максимизации правдоподобия эквивалентна минимизации взвешенной суммы KL-дивергенций с весами n_d между эмпирическими распределениями $\hat{p}(w\,|\,d) = \frac{n_{dw}}{n_d}$ и модельными распределениями $p(w\,|\,d) = \langle \varphi_w, \theta_d \rangle$.

При $\ell(p) = p^\lambda$ задача эквивалентна минимизации взвешенной суммы дивергенций Кресси–Рида (стр. 116).

При $\ell(p) = p$ задача переходит в максимизацию скалярных произведений:

$$\sum_{d\in D} n_d \big\langle \hat{p}(w\,|\,d), \langle \varphi_w, \theta_d \rangle \big\rangle + R(\Phi, \Theta) \;\to\; \max_{\Phi,\Theta}.$$

Данный принцип оптимизации модели представляется не менее разумным, чем максимизация правдоподобия. При этом $p_{tdw} = \varphi_{wt}\theta_{td}$, то есть из обычной формулы E-шага уходит знаменатель с нормировочным множителем. Будем называть это *быстрым E-шагом*. Он может давать заметное ускорение EM-алгоритма, поскольку нормировочные множители требуют вычисления скалярных произведений $\langle \varphi_w, \theta_d \rangle$.

По результатам сравнения различных стратегий оптимизации в [27] была предложена комбинированная стратегия, когда первая половина итераций использует быстрый E-шаг, при этом модель становится разреженной благодаря регуляризации. Финальная доводка модели происходит уже по максимуму правдоподобия, но теперь скалярные произведения $\langle \varphi_w, \theta_d \rangle$ могут вычисляться намного быстрее благодаря разреженности распределений φ_{wt} и θ_{td}. Этот вариант алгоритма даёт выигрыш в скорости вычислений до 30% при большом числе тем (500 и более).

Выводы по главе

- В главе описаны варианты и особенности реализации EM-алгоритма в библиотеке тематического моделирования `BigARTM`.
- Пакетный алгоритм позволяет обрабатывать большие данные порциями, не загружая текстовую коллекцию целиком в оперативную память.
- Параллельный алгоритм позволяет обрабатывать несколько пакетов одновременно, сокращая общее время обработки.
- Онлайновый алгоритм считается наиболее эффективным способом организации вычислений. Частые обновления матрицы Φ позволяют тематизировать большую коллекцию за один линейный проход по всем документам.
- Быстрый E-шаг без нормировки позволяет ускорить первые итерации алгоритма, когда высокая точность ещё не нужна.
- В следующей главе даются начальные сведения по практическому применению библиотеки `BigARTM`.

20 Проект BigARTM

BigARTM — это библиотека с открытым кодом, основанная на теории ARTM. Она имеет расширяемый встроенный набор регуляризаторов и метрик качества, реализует онлайновый и оффлайновый многопоточный пакетный EM-алгоритм, обеспечивающий высокую эффективность обработки больших коллекций на одном компьютере.

Библиотека является кроссплатформенной: сборку и исполнение можно производить под Windows 7/8/10, Mac OS и различными дистрибутивами Linux. Поддерживаются программные интерфейсы под Python 2.7.*/3.*, C++, а также запуск в виде исполняемого бинарного файла. Исходный код BigARTM написан на C++11. Поддерживается несколько популярных форматов текстовых данных.

Исчерпывающую информацию по библиотеке можно найти в документации на сайте http://bigartm.org.

Ниже представлен минимальный код в Python, выполняющий загрузку и преобразование данных во внутренний формат пакетов документов (*батчей*), создание и обучение модели, вычисление и вывод перплексии.

```
1  # Import all necessary tools and data
2  from sklearn.feature_extraction.text import CountVectorizer
3  from sklearn.datasets import fetch_20newsgroups
4  from numpy import array
5  import artm
6  # Extract data using sklearn and numpy
7  cv = CountVectorizer(max_features=1000, stop_words='english')
8  n_wd = array(cv.fit_transform(fetch_20newsgroups().data).todense()).T
9  vocabulary = cv.get_feature_names()
10 # Create batches and dictionary
11 bv = artm.BatchVectorizer(data_format='bow_n_wd',
12                            n_wd=n_wd,
13                            vocabulary=vocabulary)
14 # Learn simple PLSA model
15 model = artm.ARTM(num_topics=15, dictionary=bv.dictionary)
16 model.scores.add(artm.PerplexityScore(name='perp',
17                                        dictionary=bv.dictionary))
18 model.fit_offline(bv, num_collection_passes=20)
19 # Print perplexity values by iterations
20 print(model.score_tracker['perp'].value)
```

Подготовка данных. Универсальным объектом, принимаемым на вход всеми операциями BigARTM, является векторизатор artm.BatchVectorizer. В примере выше (шаги 10–13) он был создан по матрице «мешка слов» n_wd и словаря, задающего соответствие между строками матрицы и словами коллекции. В этом случае пакеты создаются в оперативной памяти и полностью удаляются из неё по завершении работы библиотеки. Этот способ хранения данных подходит только для небольших коллекций, целиком помещающихся в памяти. Во всех остальных случаях используются форматы данных, предполагающие чтение исходных документов с диска и запись итоговых пакетов на диск. Наиболее популярен формат текстовых файлов Vowpal Wabbit, в котором каждая строка соответствует одному документу и имеет вид

doc_title token_1:value_1 token_2:value_2 ...

Данный формат позволяет представлять документы как «мешком слов», так и последовательным текстом, а также записывать в документы термы различных модальностей. Пример создания векторизатора по данным в формате Vowpal Wabbit:

```
bv = artm.BatchVectorizer(data_path='docword.vw.txt',
                          data_format='vowpal_wabbit',
                          target_folder='my_collection_batches')
```

Здесь `data_path` — путь к файлу с документами, параметр `target_folder` указывает на несуществующую директорию для сохранения готовых пакетов.

Парсинг большой коллекции — относительно длительный процесс (даже несмотря на то, что `BigARTM` умеет выполнять его в многопоточном режиме), поэтому удобно сохранить пакеты на диск и использовать их многократно. Пример создания векторизатора, загружающего пакеты с диска:

```
bv = artm.BatchVectorizer(data_path='my_collection_batches',
                          data_format='batches')
```

Словари `BigARTM` предназначены для хранения данных о словах и используются в некоторых регуляризаторах и метриках качества. Словарю соответствует объект `artm.Dictionary`, который можно либо сформировать автоматически во время разбиения коллекции на пакеты (задав в `artm.BatchVectorizer` параметр `gather_dictionary`, по умолчанию равный `True`), либо создать вручную на основе своих данных. Объект словаря можно сохранить в бинарный или текстовый файл, затем загружать его из этого файла:

```
bv = artm.BatchVectorizer(data_path='docword.vw.txt',
                          data_format='vowpal_wabbit',
                          target_folder='my_collection_batches')
bv.dictionary.save('my_collection_batches/dictionary')
# Load dictionary back during next BigARTM launch:
dictionary = artm.Dictionary()
dictionary.load('my_collection_batches/dictionary.dict')
```

Готовый словарь можно изменять. Для этого достаточно сохранить его на диск в текстовом виде, затем модифицировать полученный файл и загрузить его обратно:

```
dictionary.save_text('my_collection_batches/dictionary.txt')
# Change file according to your needs ...
# Then, load it back
dictionary.load_text('my_collection_batches/dictionary.txt')
```

Словарь, сохранённый в текстовом файле, состоит из строк следующего вида:

token class_id value tf df

где token — строковое представление слова, class_id — модальность, tf — абсолютная частота слова в коллекции, df — число документов коллекции, в которых слово встретилось хотя бы раз. Поле value по умолчанию заполняется нормированным значением tf, но может быть изменено (эта возможность используется как механизм передачи данных в некоторых регуляризаторах и метриках качества).

Словари можно фильтровать встроенными средствами по значениям tf и df, например, можно отбрасывать слишком частые или слишком редкие слова. Указание словаря в конструкторе `artm.ARTM` или в методе `artm.ARTM.initialize` задаёт порядок строк в матрицах Φ и (n_{wt}) согласно порядку слов в словаре. Это важно, например, для модальности меток времени.

Регуляризаторы могут воздействовать на матрицы Φ, Θ или (p_{tdw}). Регуляризаторы Φ могут воздействовать на отдельные модальности. Наличие параметра `class_id` указывает, что регуляризатор работает с одной модальностью, по-умолчанию с `@default_class`. Наличие параметра `class_ids` указывает, что регуляризатор работает со списком модальностей, по-умолчанию со всеми. Почти все параметры всех регуляризаторов можно менять между итерациями обучения.

`SmoothSparsePhiRegularizer` — *регуляризатор сглаживания–разреживания матрицы* Φ, реализован по формуле (25) с небольшим обобщением:

$$\varphi_{wt} = \underset{w \in W}{\operatorname{norm}}\big(n_{wt} + \tau\beta_w f(\varphi_{wt})\big).$$

Если в определении KL-дивергенции заменить логарифм $\ln x$ на функцию $\lambda(x)$, то $f(x) = x\lambda'(x)$. По умолчанию $f(x) = 1$, что и соответствует логарифму. В библиотеке можно задавать $f(x)$ как степенную функцию. Вектор (β_w) загружается из словаря и задаётся значениями поля value каждого слова. Для каждой темы t может быть задан свой такой вектор. Таким способом можно задавать «белые» и «чёрные» списки слов для частичного обучения.

`SpecifiedSparsePhiRegularizer` — *регуляризатор разреживания* Φ, реализован по той же формуле, но $\tau\beta_w$ является константой и подбирается таким образом, чтобы доля нулевых элементов в матрице Φ оказалась не ниже заданного порога. При этом функция f не используется, то есть $f(x) = 1$.

`SmoothSparseThetaRegularizer` — *регуляризатор сглаживания–разреживания матрицы* Θ, реализован по формуле (26), с аналогичным обобщением:

$$\theta_{td} = \underset{t \in T}{\operatorname{norm}}\big(n_{td} + \tau\alpha_i\alpha_{td} f(\theta_{td})\big).$$

Функция f играет ту же роль, что и в регуляризаторе сглаживания–разреживания Φ. Массив множителей α_i позволяет управлять воздействием регуляризатора на каждой i-й внутренней итерации обработки документа. Вектор или матрица (α_{td}) позволяет управлять воздействием регуляризатора на элементы матрицы Θ.

`DecorrelatorPhiRegularizer` — *регуляризатор декоррелирования тем в матрице* Φ, реализован согласно (37). От пользователя требуется указать коэффициент регуляризации τ и список модальностей, на которые нужно воздействовать.

`TopicSelectionThetaRegularizer` — *регуляризатор отбора тем в матрице* Θ, реализован по формуле (64). Единственное отличие заключается в наличии массива множителей α_i, как в регуляризаторе сглаживания–разреживания Θ.

`SmoothTimeInTopicsPhiRegularizer` — *регуляризатор сглаживания тем по модальности времени в матрице* Φ, реализован по формуле (51). Для корректной работы регуляризатора требуется указать имя модальности времени и расположить термы времени в словаре и в матрице Φ в хронологическом порядке.

`NetPlsaPhiRegularizer` — *регуляризатор NetPLSA для модальности вершин графа в матрице* Φ, определяется по формуле (63). В документах должны быть заранее записаны термы вершин графа v. В параметрах регуляризатора задаются имена вершин v, их веса (мощности множеств $|D_v|$) и веса рёбер графа w_{uv}.

`ImproveCoherencePhiRegularizer` — *регуляризатор когерентности*, реализован по формуле (70) и в качестве параметра требует словарь парной сочетаемости слов C_{wv} (собрать его можно с помощью встроенного парсера).

`BitermsPhiRegularizer` — *регуляризатор битермов*, реализован по формулам (67)–(68) и в качестве параметра также требует словарь частот битермов n_{uv} (задача его сборки ложится на пользователя).

`LabelRegularizationPhiRegularizer` — *частотный регуляризатор матрицы* Φ для классификации с несбалансированными классами. Реализован по формуле (60). В качестве параметра требует словарь классов со значениями их мощностей $|D_c|$.

`HierarchySparsingThetaRegularizer` — *регуляризатор иерархического разреживания* Θ, используется для разреживания матрицы связей между родительскими темами и их дочерними подтемами в иерархических моделях, согласно формуле (66).

`TopicSegmentationPtdwRegularizer` — *регуляризатор E-шага для разреживания сегментов в матрицах* (p_{tdw}), определяемый по формуле (93).

`SmoothPtdwRegularizer` — *регуляризатор E-шага для сглаживания матриц* (p_{tdw}) по локальному контексту. Приближает тематический профиль каждого вхождения терма к усредненному профилю его соседей (по окну фиксированной ширины).

Регуляризаторы могут включаться, отключаться или модифицироваться в любой момент между вызовами `fit_offline` или `fit_online`, что позволяет, в совокупности с контролем метрик качества, гибко перестраивать стратегию регуляризации в соответствии с текущим состоянием модели. Пример:

```
reg = artm.DecorrelatorPhiRegularizer(name='decor', tau=1e+5)
model.regularizer.add(reg)
model.scores.SparsityPhiScore(name='sparse')

model.fit_offline(batch_vectorizer=bv, num_collection_passes=10)
print model.score_tracker('sparse').last_value

# Printing result: 0.15 - too small. Let's increase tau
model.regularizer['decor'].tau = 3e+5
model.fit_offline(batch_vectorizer=bv, num_collection_passes=15)
```

Многопоточный пакетный EM-алгоритм. Библиотека `BigARTM` позволяет обрабатывать коллекции документов, не помещающиеся в оперативную память. Для этого коллекция D с помощью `BatchVectorizer` разбивается на пакеты D_b, $b = 1, \ldots, B$, каждый из которых хранится в отдельном файле. Обычно используются пакеты размером от сотен килобайт до десятков мегабайт. Коэффициенты регуляризации задаются в момент создания модели, но потом могут быть в любой момент изменены, в том числе в ходе EM-итераций.

Пакеты обрабатываются функцией ProcessBatches как описано в алгоритме 6. Оффлайновый алгоритм FitOffline запускается функцией `ARTM.fit_offline`, онлайновый FitOnline — функцией `ARTM.fit_online`. Для включения асинхронного алгоритма последней надо передать параметр `async=True`. Контроль условий сходимости EM-алгоритма возлагается на пользователя. Проще всего задавать число итераций по коллекции и по каждому документу.

Построенную тематическую модель можно использовать для тематизации отдельных документов при фиксированной матрице Φ. Эта возможность реализуется функцией `ARTM.transform`, которая пропускает документы через ProcessBatches.

Метрики качества добавляются через поля `scores` объекта `ARTM`. В этот момент у многих метрик можно задавать параметры. Вычисленные значения метрик извлекаются через поля `score_tracker`.

`PerplexityScore` — *перплексия*, вычисляемая по формуле (95). Для её корректной работы нужен словарь, содержащий нормированные частоты слов в коллекции (не модифицированные значения value для каждого слова). Они используются в качестве аппроксимации нулевых значений $p(w\,|\,d)$ и позволяют корректно оценивать модели на одном словаре, но с разной степенью разреженности.

Пример подключения перплексии для модальности `@default_class` (стандартная модальность слов):

```
# m = artm.ARTM(...)
m.scores.add(artm.PerplexityScore(name='perp',
                                  class_ids=['@default_class'],
                                  dictionary=dictionary))
```

Значения перплексии можно вывести следующим образом (вместо value можно вывести, например, числитель и знаменатель перплексии по каждой модальности):

```
print(model.score_tracker['perp'].value)
```

Поле value содержит всю историю значений метрики по обновлениям матрицы Φ. У любого поля любой метрики имеется вариант с префиксом `last_`, который возвращает значение метрики на момент последней синхронизации. Это может быть полезно для получения массивных метрик типа `TopTokensScore`.

`SparsityPhiScore/SparsityThetaScore` — *разреженности матриц* Φ *и* Θ. Оцениваются долей элементов матрицы, меньших заданного пользователем порога.

`TopTokensScore` — *топ-слова в темах*, список из заданного числа слов с наибольшей вероятностью по каждой теме. Если в параметрах этой метрики указать словарь парной сочетаемости слов, то будет вычислена когерентность coher_t по спискам топ-слов в темах, согласно формуле (96).

`TopicKernelScore` — *ядровые характеристики тем:* чистота pur_t, контрастность con_t, размер ядра ker_t оценивающие различность и, косвенно, интерпретируемость каждой темы t, см. стр. 111. Аналогично топ-словам, указание словаря парной сочетаемости запускает подсчёт когерентности, но теперь уже по ядрам тем.

`BackgroundTokensRatioScore` — *доля фоновых слов*, оценивает долю слов, для которых KL-дивергенция между распределениями $p(t)$ и $p(w\,|\,t)$ выше заданного порога.

`TopicMassPhiScore` — *частоты тем* n_t и распределения $p(t) = \frac{n_t}{n}$ для всех тем t, вычисляемые по матрице (n_{wt}).

`ItemsProcessedScore` — *число обработанных документов*, техническая метрика, показывающая по итерациям количество документов (с повторами), обработанных EM-алгоритмом с момента включения метрики.

`PeakMemoryUsage` — *пиковое потребление памяти*, техническая метрика (доступная только в C++ интерфейсе), предоставляющая информацию о максимальном потреблении оперативной памяти за время каждой итерации алгоритма.

Пользователь может не только создавать собственные метрики, но и вычислять их напрямую, сделав выгрузку параметров модели Φ и Θ.

Выгрузка параметров модели. В следующем коде показано, как получить матрицу Φ (точнее, первые 10 тем дефолтной модальности):

```
model.get_phi(topic_names=model.topic_names[: 10],
              class_ids=['@default_class'],
              model_name=model.model_pwt)
```

Указание параметра `model_name=model.model_nwt` позволяет аналогичным образом получить значения n_{wt} вместо φ_{wt}.

С матрицей Θ можно работать по-разному. Во-первых, её можно вообще не хранить, если она не нужна. Во-вторых, можно хранить её в кэше, задав перед началом обучения параметр `cache_theta=True`. В третьих, можно включить хранение Θ в Φ-подобной матрице, задав параметр `theta_name`. Это даст свободный доступ к матрице на чтение и запись в любой момент. В первом случае выгрузить матрицу невозможно, в остальных применим следующий код:

```
# case 2
model.get_theta()
# case 3
model.get_phi(model_name=model.theta_name)
```

Все описанные вызовы возвращают объекты `pandas.DataFrame`.

Помимо описанного интерфейса выгрузки матриц, есть возможность получить указатель на матрицу и напрямую модифицировать память, используемую ядром библиотеки, что существенно уменьшает расход памяти и время вычислений.

Выводы по главе

- В главе даны начальные сведения по практическому применению библиотеки `BigARTM`, включая использование встроенных регуляризаторов и метрик качества.
- Более подробную информацию и примеры кода на языке Python можно найти в документации на сайте `http://bigartm.org`.

21 Разведочный поиск и другие приложения

Важным приложением тематического моделирования является *информационный поиск* (information retrieval) [203, 23]. Современные поисковые системы предназначены, главным образом, для поиска конкретных ответов на короткие текстовые запросы. Совсем другие поисковые потребности возникают у пользователей, которым необходимо разобраться в новой предметной области или пополнить свой багаж знаний. Пользователь может не владеть терминологией, слабо понимать структуру предметной области, не иметь точных формулировок запроса и не подразумевать единственный правильный ответ. В таких случаях нужен поиск не по ключевым словам, а по смыслу. Запросом может быть длинный фрагмент текста, документ или подборка документов. Результатом поиска может быть обычная поисковая выдача в виде списка документов или более структурированное представление с разделением по темам или графическая визуализация — «дорожная карта» предметной области, статическая или интерактивная.

Для этих случаев подходит парадигма *разведочного информационного поиска* (exploratory search) [112, 199]. Его целью является получение ответов на сложные вопросы: «какие темы представлены в тексте запроса», «что читать в первую очередь по этим темам», «что находится на стыке этих тем со смежными областями», «какова тематическая структура данной предметной области», «как она развивалась во времени», «каковы последние достижения», «где находятся основные центры компетентности», «кто является экспертом по данной теме» и т. д. Пользователь обычной поисковой системы вынужден итеративно переформулировать свои короткие запросы, расширяя зону поиска по мере усвоения терминологии предметной области, периодически пересматривая и систематизируя результаты поиска [139]. Это требует затрат времени и высокой квалификации. Разведочный поиск реализуется с помощью визуальных тексто-графических представлений «общей картины», дающих пользователю больше уверенности, что все важнейшие аспекты изучаемой проблемы найдены и отображены. Если образно представить итеративный поиск как блуждание по лабиринту знаний с фонариком, то разведочный поиск — это автоматическое построение карты для любой части этого лабиринта.

Тематический поиск. Полнотекстовые поисковые системы основаны на инвертированных индексах, в которых для каждого слова хранится список содержащих его документов [15]. Поисковая система ищет документы, содержащие все слова запроса, поэтому по длинному запросу, скорее всего, ничего не будет найдено.

Система тематического разведочного поиска сначала строит тематическую модель запроса и определяет короткий список тем запроса. Затем для поиска документов схожей тематики применяются те же механизмы индексирования и поиска, только в роли слов выступают темы. Поскольку число тем на несколько порядков меньше объёма словаря, тематический поиск требует намного меньше памяти по сравнению с полнотекстовым поиском и может быть реализован на более скромной технике.

Технологии информационного поиска на основе тематического моделирования долгое время находились в стадии исследований [168, 35, 142, 43, 21, 190], но так и не вышли на уровень широкого коммерческого применения. В литературе по разведочному поиску на тематическое моделирование иногда ссылаются [157, 71, 154, 177], однако многие обзоры о нём вообще не упоминают [66, 150, 166, 87, 113, 83]. С другой

стороны, в работах по тематическому моделированию разведочный поиск часто называют одним из важнейших приложений, а оценки качества поиска используют для валидации тематических моделей [203, 23]. Это говорит как о разобщённости исследовательских сообществ, так и о том, что тематическое моделирование пока находит лишь нишевые применения в разведочном информационном поиске.

В статье [177] важными преимуществами тематических моделей для поиска называются гибкость, возможности визуализации и навигации. В качестве недостатков отмечаются проблемы с интерпретируемостью тем, трудности с модификацией тематической модели при поступлении новых документов и высокая вычислительная сложность. Однако эти проблемы относятся к устаревшим методам и были успешно решены в последующие годы: десятки новых моделей разработаны для улучшения интерпретируемости; онлайновые алгоритмы способны обрабатывать большие коллекции и потоки документов за линейное время [122, 33, 178].

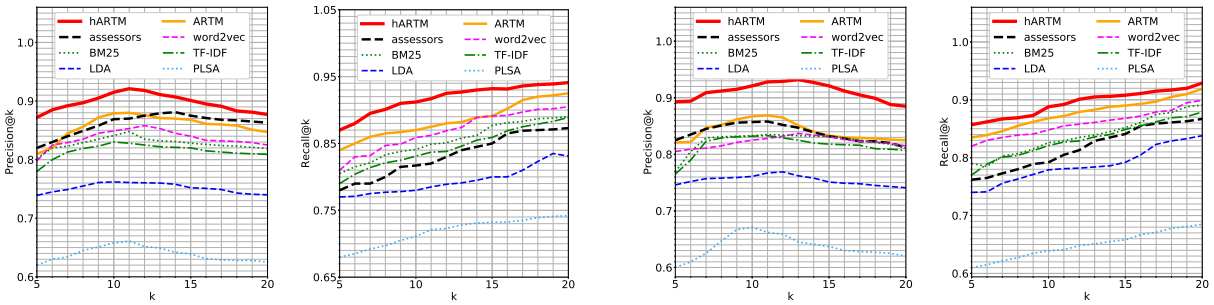
Оценивание качества тематического разведочного поиска. Модель ARTM для разведочного поиска была предложена в [20] и улучшена в [80, 81, 82]. Для измерения качества разведочного тематического поиска использовались критерии точности и полноты на основе оценок асессоров.

Была составлена выборка из 100 запросов — творческих заданий разведочного поиска. Каждый запрос представлял собой текст объёмом около одной страницы формата A4, описывающий тематику поиска. Каждое задание сначала выполнялось независимо несколькими асессорами, затем системой тематического поиска, затем её результат снова оценивался асессорами. Данная методика позволяет, единожды сделав разметку результатов поиска, многократно оценивать качество различных тематических моделей и механизмов поиска.

Эксперименты по подбору тематической модели для разведочного поиска проводились на двух коллекциях: 175 тысяч статей русскоязычного коллективного блога `habrahabr.ru` и 760 тысяч статей англоязычного блога `techcrunch.com`.

Было показано, что тематический поиск находит больше релевантных документов, чем асессоры, которые затрачивают на выполнение таких заданий 30 ± 20 минут. Комбинирование регуляризаторов декоррелирования, разреживания и сглаживания вместе с модальностями n-грамм, авторов и категорий значимо улучшает качество поиска (см. рис. 26) и позволяет достичь точности и полноты 85% и выше при тщательном подборе числа тем $|T|$, см. рис. 27.

Применение иерархических тематических моделей в [81] позволило улучшить точность и полноту ещё на 5–8% при одновременном увеличении оптимального числа тем $|T|$ в несколько раз, см. рис. 28. Этот результат можно интерпретировать следующим образом: постепенное дробление тем на подтемы способствует более аккуратному разреживанию тематических векторов документов $p(t\,|\,d)$. Темы, которых точно нет в данном документе, получают нулевую вероятность уже на первом–втором уровне иерархии. Поэтому мелко гранулированные подтемы нижнего уровня, которые могли бы быть статистически ненадёжными, наследуют нулевые вероятности от своих родительских тем. Как результат, поисковая выдача очищается от нерелевантных «мусорных» документов, что и означает увеличение именно точности поиска. Полнота поиска также немного улучшается, возможно, благодаря увеличению размерности тематических векторов. Можно сказать и так, что иерархические темати-



(а) качество поиска по коллекции Habrahabr.ru

(б) качество поиска по коллекции TechCrunch.com

Рис. 26: Точность (Precision@k) и полнота (Recall@k) по первым k позициям выдачи при поиске по длинным запросам. Документы ранжировались по косинусной мере сходства векторных представлений. Иерархическая модель hARTM опережает по точности и полноте плоскую модель ARTM, при этом обе модели регуляризованные (декоррелирование и сглаживание матрицы Φ, разреживание матрицы Θ) и мультимодальные (n-граммы, авторы, категории). Модели ранжирования, основанные на сходстве векторов слов (TF-IDF, BM25) или нейросетевых эмбедингов (word2vec), являются сильными конкурентами, немного опережая людей (assessors) по полноте, но отставая от них по точности. Классические тематические модели PLSA и LDA не выдерживают конкуренции [81].

	Habrahabr.ru						TechCrunch.com					
	acecc	100	150	**200**	250	400	acecc	350	400	450	**475**	500
P@5	0.821	0.662	0.721	**0.810**	0.761	0.693	0.822	0.653	0.725	0.752	**0.819**	0.777
P@10	0.869	0.761	0.812	**0.879**	0.825	0.673	0.851	0.663	0.732	0.762	**0.867**	0.811
P@15	0.875	0.733	0.795	**0.868**	0.791	0.651	0.835	0.682	0.743	0.787	**0.833**	0.793
P@20	0.863	0.724	0.795	**0.847**	0.792	0.642	0.813	0.650	0.743	0.773	**0.825**	0.793
R@5	0.780	0.732	0.807	**0.840**	0.821	0.721	0.762	0.731	0.762	0.793	**0.835**	0.817
R@10	0.817	0.771	0.843	**0.870**	0.851	0.751	0.792	0.763	0.793	0.812	**0.868**	0.855
R@15	0.850	0.824	**0.895**	0.891	0.871	0.773	0.835	0.782	0.807	0.855	**0.890**	0.882
R@20	0.873	0.857	0.905	**0.925**	0.892	0.771	0.867	0.792	0.823	0.862	**0.919**	0.903

Рис. 27: Подбор оптимального числа тем по критериям точности (P@k) и полноты (R@k) поиска для «плоской» модели ARTM с регуляризаторами (декоррелирование и сглаживание матрицы Φ, разреживание матрицы Θ) и модальностями (n-граммы, авторы, категории) на коллекциях Habrahabr.ru и TechCrunch.com. Для сравнения показаны оценки качества поиска, выполненного асессорами.

ческие представления документов лучше соответствует человеческой логике поиска: когда мы ищем информацию, мы сразу отсекаем то, что нам точно не подходит.

Абляционные эксперименты в [81] показали, что все регуляризаторы и модальности важны: выключение каждого из них из модели приводит к падению точности и полноты поиска.

Тематические модели в социо-гуманитарных исследованиях. Известно множество прикладных задач и сценариев применения тематического моделирования в цифровых гуманитарных исследованиях (digital humanities), филологии, культурологии, социологии, политологии, истории [44, 198, 86].

Чаще всего исследование начинается с вопроса «как понять, о чём все эти тексты, не читая их» или «как выделить из них ту часть, которая представляет интерес для более пристального изучения». При этом исследователи могут не знать заранее, какие именно темы или аспекты окажутся интересными и приведут к новым знаниям или нетривиальным выводам. По сути, это также разновидность разведочного информационного поиска.

Habrahabr.ru

$\|T_1\|$	**20**		**25**							**30**	
$\|T_2\|$	150	200	**250**		**275**			**300**		400	450
P@5	0.621	0.742	0.839	0.850	0.865	**0.869**	**0.869**	0.803	0.769	0.701	0.670
P@10	0.645	0.749	0.850	0.861	0.879	**0.911**	0.895	0.809	0.796	0.719	0.689
P@15	0.635	0.751	0.848	0.869	0.873	**0.893**	0.887	0.807	0.781	0.721	0.701
P@20	0.630	0.745	0.841	0.855	0.864	0.874	**0.875**	0.800	0.775	0.709	0.675
R@5	0.628	0.773	0.843	0.865	0.881	**0.881**	0.868	0.849	0.839	0.715	0.691
R@10	0.652	0.782	0.855	0.871	0.902	**0.918**	0.877	0.871	0.845	0.745	0.699
R@15	0.671	0.801	0.870	0.889	0.929	**0.939**	0.901	0.883	0.861	0.781	0.722
R@20	0.680	0.819	0.886	0.892	**0.955**	**0.955**	0.907	0.901	0.872	0.801	0.729
$\|T_3\|$	750	800	1200	1300	1300	**1400**	1500	1500	1600	3000	3500
P@5	0.625	0.743	0.840	0.852	0.869	**0.872**	0.870	0.805	0.771	0.705	0.672
P@10	0.648	0.754	0.851	0.867	0.882	**0.915**	0.901	0.811	0.799	0.722	0.694
P@15	0.632	0.752	0.850	0.872	0.878	**0.895**	0.889	0.809	0.785	0.729	0.703
P@20	0.629	0.745	0.845	0.861	0.871	0.877	**0.882**	0.803	0.778	0.710	0.681
R@5	0.632	0.780	0.845	0.869	0.883	**0.889**	0.872	0.851	0.841	0.721	0.695
R@10	0.654	0.792	0.859	0.873	0.905	**0.922**	0.881	0.873	0.850	0.749	0.703
R@15	0.675	0.805	0.874	0.892	0.932	**0.942**	0.905	0.889	0.863	0.787	0.725
R@20	0.684	0.824	0.889	0.901	0.958	**0.961**	0.912	0.904	0.878	0.805	0.734

TechCrunch.com

$\|T_1\|$	**80**		**100**							**120**	
$\|T_2\|$	300	350	**500**		**550**			**600**		700	750
P@5	0.651	0.701	0.749	0.789	0.883	**0.889**	**0.889**	0.785	0.721	0.701	0.675
P@10	0.675	0.709	0.771	0.821	0.891	**0.918**	0.902	0.803	0.738	0.718	0.691
P@15	0.687	0.712	0.773	0.827	0.899	**0.919**	0.905	0.817	0.741	0.721	0.701
P@20	0.683	0.707	0.759	0.815	0.885	0.888	**0.895**	0.805	0.732	0.716	0.679
R@5	0.749	0.791	0.801	0.854	0.868	**0.875**	0.861	0.849	0.829	0.731	0.701
R@10	0.765	0.809	0.823	0.873	0.890	**0.904**	0.875	0.867	0.835	0.745	0.708
R@15	0.771	0.820	0.841	0.882	0.909	**0.921**	0.895	0.890	0.848	0.769	0.717
R@20	0.778	0.825	0.851	0.887	0.928	**0.942**	0.929	0.901	0.869	0.785	0.728
$\|T_3\|$	1500	1700	2500	2600	2600	**2800**	3000	3000	3200	4500	4700
P@5	0.655	0.707	0.751	0.792	0.887	**0.893**	0.890	0.789	0.722	0.703	0.678
P@10	0.678	0.712	0.773	0.823	0.895	**0.922**	0.905	0.805	0.741	0.722	0.692
P@15	0.692	0.715	0.775	0.831	0.902	**0.921**	0.907	0.821	0.743	0.725	0.703
P@20	0.687	0.709	0.761	0.819	0.889	0.885	**0.898**	0.809	0.736	0.719	0.683
R@5	0.751	0.795	0.802	0.856	0.871	**0.877**	0.863	0.852	0.831	0.738	0.705
R@10	0.767	0.812	0.825	0.875	0.892	**0.908**	0.879	0.871	0.842	0.751	0.711
R@15	0.772	0.824	0.841	0.887	0.912	**0.927**	0.901	0.893	0.854	0.772	0.721
R@20	0.783	0.830	0.854	0.892	0.931	**0.949**	0.935	0.905	0.871	0.790	0.732

Рис. 28: Подбор оптимального числа тем по критериям точности (P@k) и полноты (R@k) в двухуровневой и трёхуровневой иерархической модели ARTM с регуляризаторами (декоррелирование и сглаживание матрицы Φ, разреживание матрицы Θ) и модальностями (n-граммы, авторы, категории). Для каждого уровня ℓ подобрано своё оптимальное число тем $|T_\ell|$. Трёхуровневая модель достигает лучшего качества поиска в сравнении с двухуровневой или «плоской» моделью. При этом значительно увеличивается оптимальное число тем на нижнем уровне.

Типичный сценарий заключается в том, чтобы задать поисковую тематику с помощью словаря ключевых слов или подборки текстов. Эту задачу мы уже рассматривали, когда говорили про *сфокусированный тематический поиск*, стр. 50.

Среди найденных тем могут оказаться как хорошие (интерпретируемые), так и мусорные, а среди хороших — как релевантные, так и нерелевантные. Получив тематическую модель, пользователь может разметить темы на эти три типа, и затем при повторном перестроении модели потребовать, чтобы релевантные темы сохранились, а на месте мусорных, по возможности, возникли новые хорошие темы. При этом данные предыдущей модели и пользовательской разметки используются для конструирования регуляризаторов сглаживания и декоррелирования.

Выделив интересующие темы, пользователь может построить следующую модель, в которой попытаться разделить эти темы на большее количество мелко гранулированных подтем.

Многие исследования, проводимые на данных социальных сетей, новостных потоков, старых газетных и журнальных подборок, исторических архивов связаны с выявлением и анализом динамики процессов во времени. Исследователей могут интересовать цепочки событий, появление новых аспектов, изменения в употреблении слов, связанные с культурными влияниями и сдвигами.

Приведём несколько примеров таких исследований.

В [126] изучалась полная коллекция публикаций 24 научных журналов по филологии и археологии за полтора столетия (1850–2006 гг.), доступная через цифровую базу данных JStor. Прослеживалась динамика изменений в тематике и терминологии. JStor предоставляет тексты в формате «мешков слов» из-за ограничений прав собственности; в этих условиях именно тематическое моделирование оказалось адекватным инструментом анализа.

В [70] изучалась научная периодика в области литературы, охватывающая более 21 тысячи статей за период 1889–2013 гг. Особое внимание уделялось изменениям тем и значений слов внутри каждой темы, в частности, трансформации отношения к теме насилия.

В [114] изучалась коллекция газет и периодики Финляндии за период 1854–1917 гг. Целью было выделение тем о церкви, религии и образовании для понимания тематических трендов модернизации и секуляризации финского общества.

В [158] изучалась коллекция выступлений в Совбезе ООН по Афганистану за период 2001–2017 гг. Решалась задача выявления динамики отношения разных стран к проблемам Афганистана.

В [63] изучалась коллекция 7,6 тысяч новостных статей в СМИ Пакистана за период 2010–2021 гг. Решалась задача выявления тем, связанных с изменениями климата, выявленные темы вручную группировались по научным, социальным и политическим аспектам климатической повестки; исследовалась динамика тем во времени.

В [74] изучалась коллекция AYLIEN COVID-19, содержащая 1,5 млн. новостей о пандемии из 440 источников СМИ за период 11.2019–07.2020. Решалась задача выявления тем, вызывающих поляризацию общественного мнения, связанную с политическими пристрастиями или партийностью.

Несмотря на большое количество разнообразных применений в цифровых гуманитарных исследованиях, тематические модели критикуются за несоответствие статистической и лингвистической концепции «темы», что порождает ряд эпистемологических проблем [159]. Более того, делается вывод, что тематическое моделирование

в его нынешнем состоянии не отвечает требованиям методологической интеграции с признанными методами анализа и не может использоваться как самостоятельный метод исследования. «Тематическая модель исходит из относительно нереалистичных предположений, имеет недетерминированный результат, не может быть эффективно проверена в сравнении с разумным числом конкурирующих моделей, не привязана к четко определенному лингвистическому интерфейсу. Например, статистические темы (topics) не соответствуют содержательным темам (themes) в контент-анализе. Из-за этих особенностей интерпретация модели оказывается подверженной апофении (человеческая склонность воспринимать случайные наборы элементов как значимые паттерны) и предвзятости подтверждения (человеческая склонность предпочитать паттерны, соответствующие априорным предубеждениям). В то время как частичная проверка статистической модели возможна, её концептуальная проверка требует широкого взаимодействия с другими методами и человеческими оценками, а также выяснения того, коррелирует ли лексическая сочетаемость с концептуальными темами» [159].

Вполне оправданная критика во многом связана с тем, что в тематическом моделировании предпринималось недостаточно усилий для формализации лингвистических знаний и требований. В ряде исследований тематическое моделирование используется как вспомогательный инструмент контент-анализа [52, 72, 22, 31, 30], при этом используется морально устаревшая модель LDA, а задачи модификации модели под потребности лингвистического анализа даже не ставятся.

Развитие тематического моделирования до 2016 года было направлено в значительной степени на решение внутренних математических проблем, связанных с применением байесовского вывода [51]. Дальнейшее развитие направилось в сторону интеграции с нейросетевыми моделями языка и снова сосредоточилось на преодолении трудностей технического характера [209].

Решение концептуальных проблем тематического моделирования требует глубокой междисциплинарной экспертизы, тогда как поверхностный «технократический» подход к цифровым гуманитарным исследованиям обречён на неудачу. Теория ARTM является определённым шагом в этом направлении, но не готовым решением всех проблем, а гибкой теоретической основой, способной снимать существовавшие ранее барьеры и подсказывать направления развития.

Требования к тематическим моделям. Анализ публикаций и собственных прикладных исследований показывает, что на практике к тематической модели всегда предъявляется совокупность требований. Тематическая модель для разведочного поиска или социо-гуманитарных исследований должна быть одновременно:

- *хорошо интерпретируемой* в тех случаях, когда темы используются в пользовательском интерфейсе для полуавтоматического поиска и фильтрации данных, уточнения критериев поиска, навигации по коллекции и визуализации результатов поиска;

- *разреженной*, чтобы каждый документ состоял из небольшого числа тем — это упрощает аналитику и повышает эффективность поиска;

- *иерархической*, чтобы пользователь мог получить представление о тематической структуре предметной области на любом уровне детализации; кроме того,

иерархическое деление тем на подтемы позволяет увеличить размерность тематических векторов для повышения качества поиска;

- способной автоматически *определять число тем* на каждом уровне иерархии, автоматически *создавать новые темы* и автоматически *именовать темы* релевантными заголовками;
- *мультиграммной*, так как выделение ключевых фраз и терминов существенно улучшает интерпретируемость тем;
- *мультиязычной* в тех приложениях, где требуется кросс-языковой или мультиязыковой поиск, например, при поиске патентов на одном языке по патенту на другом языке;
- *мультимодальной*, чтобы учитывать метаданные документов: авторов, источники, ссылки, категории, теги, и др.;
- *темпоральной*, чтобы выявлять динамику развития тем, обнаруживать новые темы и быстро растущие тематические тренды;
- *сегментирующей*, чтобы не только находить релевантные документы, но и указывать в них сегменты, наиболее релевантные тематике запроса;
- *обучаемой* по пользовательской разметке, чтобы детализировать тематический анализ коллекции или персонализировать алгоритмы поиска и ранжирования;
- *онлайновой*, *параллельной* и *распределённой* с точки зрения программной реализации, чтобы эффективно обрабатывать большие коллекции текстов.

Важнейшим требованием к инструментарию тематического моделирования оказывается его способность строить модели, объединяющие в себе любые из перечисленных возможностей в различных сочетаниях. Инструментарий должен быть «лего-конструктором», позволяющим собирать модели с заданным свойствами из заранее заготовленных модулей.

Теория аддитивной регуляризации тематических моделей (ARTM) даёт вполне практичный ответ на вопрос «что есть модуль». Это регуляризатор, который легко встраивается в EM-алгоритм и наделяет модель определённым свойством.

Модульный подход поддерживается библиотекой `BigARTM`. Некоторой трудностью пока остаётся подбор коэффициентов регуляризации и других гиперпараметров, что требует понимания теории, а также навыков программирования и проведения вычислительных экспериментов в Python. Создание удобной для конечного пользователя среды тематического моделирования на основе теории ARTM и библиотеки `BigARTM` ведётся в настоящее время.

Визуализация. Систематизация результатов тематического поиска невозможна без интерактивного графического представления. В обзоре [2] описываются и сравниваются 16 средств визуализации тематических моделей на основе веб-интерфейсов. Ещё больше идей можно почерпнуть из интерактивного обзора[7], который насчитывает более 440 средств визуализации текстов.

[7] `http://textvis.lnu.se` — интерактивный обзор средств визуализации текстов.

Несмотря на такое богатство технических решений, основных идей визуализации тематических моделей не так много: это либо двумерное отображение семантической близости тем в виде графа или карты, либо тематическая иерархия, либо динамика развития тем во времени, либо графовая структура взаимосвязей между понятиями, темами, документами, авторами или иными модальностями, либо сегментная тематическая структура отдельных документов [2].

Графическая визуализация больших данных практически бесполезна в статичном исполнении, но может оказаться мощным когнитивным средством в случае интерактивной реализации. Принцип интерактивного визуального поиска информации известен как *мантра Шнейдермана*: «сначала крупный план, затем масшабирование и фильтрация, детали по требованию» [164].

Графическое отображение результатов тематического моделирования и разведочного поиска согласуется с концепцией *дальнего чтения* (distant reading), предложенной социологом литературы Франко Моретти [128]. Он противопоставляет этот способ изучения текстов нашему привычному *пристальному чтению* (close reading). Невозможно прочитать миллионы книг или статей, но вполне возможно применить статистические методы и графическую визуализацию, чтобы понять в общих чертах, о чём вся эта литература, увидеть нетривиальные общие закономерности и научиться быстрее отыскивать нужное. «Дальнее чтение — это специальная форма представления знаний, в которой меньше элементов, грубее смысл их взаимосвязей, остаются лишь формы, отношения, структуры, модели» [128].

Выводы по главе

- Приложения в области разведочного информационного поиска и социо-гуманитарных исследований предъявляют нетривиальные комбинации требований к тематическим моделям.
- В теории ARTM и библиотеке BigARTM широкий класс требований возможно формализовать и комбинировать с помощью регуляризаторов, модальностей и гиперграфовых моделей.
- Несмотря на два десятилетия исследований в области тематического моделирования, многие проблемы формализации лингвистических и гуманитарных знаний до сих пор не решены.
- Открытые проблемы тематического моделирования обсудим в следующей, заключительной, главе.

22 Заключение

О замене теоретического фундамента. Вероятностное тематическое моделирование является относительно молодым направление исследований на стыке машинного обучения и автоматической обработки текстов. Ему чуть более 20 лет. За это время успела сложиться научная традиция — рассматривать эту область как часть или как приложение байесовского обучения. В этой книге предпринята попытка преодолеть эту традицию, предложив намного более простой, но не менее выразительный теоретический фундамент в виде классической не-байесовской регуляризации.

Сотни байесовских алгоритмов тематического моделирования, описанных в тысячах публикаций за последние два десятилетия, могут быть дебайесизированы и выведены в одно действие с помощью леммы о максимизации гладкой функции на единичных симплексах. Вероятностное тематическое моделирование можно теперь называть «теорией одной леммы». Объяснять, строить и комбинировать тематические модели стало на порядок проще.

Может возникнуть вопрос — почему 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 следующему общему сценарию. Сначала формализация модели и оптимизационной задачи; затем обогащение постановки задачи дополнительными структурами и критериями, в том числе с помощью регуляризации; 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. Например, для получения интервальных оценок, проверки статистических гипотез, сэмплирования моделей, оценивания их устойчивости и прочих видов вероятностного анализа.

В тематическом моделировании типовой сценарий 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 исключительно ради получения точечных оценок максимального правдоподобия. Нет никаких значимых потребностей, которые вынуждали бы использовать байесовское обучение.

Теория аддитивной регуляризации (ARTM) есть попытка восполнить этот пробел. Возможно, попытка запоздалая, поскольку фокус интереса научного сообщества уже переключился на глубокие нейросетевые модели языка, модели внимания и архитектуры трансформеров [51]. Тематическое моделирование теперь больше сосредоточено на интеграции с нейронными сетями в поисках возможностей для «объединения лучшего от двух миров» [209].

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

Оба вида моделей тяготеют к гомогенизации [138], то есть использованию единого векторного пространства для описания разнородных объектов по данным об их взаимодействиях. В главах 9 и 10 было показано, как это реализуется в мультимодальных и гиперграфовых тематических моделях.

Оба вида моделей генерируют как глобальные эмбединги, так и локальные для каждого слова в его контексте. В главе 15 было показано несколько подходов к тематическому моделированию последовательного текста.

Нейросетевые модели намного сложнее, их эмбединги способны вобрать в себя намного больше информации о связях между словами, причём мы даже не понимаем, какие именно связи и как именно учитываются.

Тематические модели намного проще, их эмбединги учитывают только лексическую сочетаемость слов, зато сохраняют интерпретируемость. Свойство покоординатной интерпретируемости является прямым следствием того, что тематические эмбединги являются дискретными вероятностными распределениями, то есть неотрицательными нормированными векторами.

Отказ от байесовского вывода тематических моделей упрощает их возможную интеграцию с нейросетевыми моделями. Любой векторный параметр нейронной сети, если на него наложить ограничения неотрицательности и нормировки, можно обучать с помощью мультипликативного градиентного шага с проекцией на единичный симплекс, согласно «основной лемме». При этом для вычисления самого градиента можно по-прежнему использовать обратное распространение ошибки. Реализация этой идеи в системах PyTorch или TensorFlow позволит применять их для обучения произвольных тематических моделей с аддитивной регуляризацией.

О мифах в тематическом моделировании. Доминирование байесовского подхода, разобщённость академического и индустриального сообщества привели к некоторым заблуждениям о тематическом моделировании, достойным упоминания и критики.

- «Тематическое моделирование — это в основном LDA». Нет, есть сотни моделей, решающих разнообразные задачи, с которыми LDA справляется хуже.
- «Тематическое моделирование подходит только для анализа текстов». Нет, есть модели для анализа изображений, видео, графов, сигналов, транзакций.
- «Тематическое моделирование — это раздел байесовского обучения». Нет, большинство моделей гораздо проще строятся в ARTM без байесовского вывода.
- «Тематические модели служат для предсказания частоты слов в документах». Формально да, но их цель в выявлении кластерной структуры коллекции.
- «Темы часто оказываются дублирующими или плохо интерпретируемыми». Это так в LDA. Проблема решается с помощью других регуляризаторов.
- «Тематические модели основаны на гипотезе мешка слов». Многие, но не все. Тематические модели n-грамм, битермов, предложений, сегментации, регуляризация и локализация E-шага учитывают порядок слов в документах.
- «Тематические векторы не так хорошо отражают смыслы слов, как word2vec». Они делают это не хуже, если строить тематическую модель по частотам парных сочетаний слов, как это и делается в word2vec.
- «Тематическая модель LDA переобучается гораздо меньше, чем PLSA» [41]. Нет, их качество примерно одинаково, особенно на больших коллекциях.

LDA осторожнее оценивает вероятности редких слов. Формально это улучшает правдоподобие, однако эти слова практически не важны для описания тем.

- «Тематическая модель LDA имеет намного меньше параметров, чем PLSA». Нет, матрицы Φ и Θ оцениваются в обеих моделях, поскольку они нужны для приложений. На самом деле в LDA больше параметров, добавляются β и α.

- «Изучение тематического моделирования начинается с вероятностных графических моделей и байесовского вывода». Не обязательно. Гораздо проще начинать с основной леммы о максимизации на единичных симплексах.

Об открытых проблемах. Не претендуя на полноту, укажем восемь проблем, которые представляются наиболее актуальными направлениями исследований в тематическом моделировании.

- *Гарантии качества тем для несбалансированных коллекций.* Модель может выдавать плохо интерпретируемые, дублирующие или мусорные темы, если в исходной коллекции темы различаются по объёму в сотни раз. Максимизация правдоподобия имеет тенденцию выравнивать темы по объёму, а не по семантической однородности. В результате крупные темы дробятся на темы-дубликаты, а мелкие объединяются и образуют мусорные темы. Проблема несбалансированности представляется основной причиной плохой интерпретируемости тем.
  Открытая проблема: сконструировать оптимизационный критерий так, чтобы модель сама определяла, какие темы являются крупными, какие мелкими в несбалансированной коллекции.

- *Обеспечение устойчивости тематических моделей.* При многократных запусках моделирования с разными регуляризаторами, из разных случайных начальных приближений или с использовании стохастического E-шага, могут получаться модели с различными наборами тем. Построение полного набора интерпретируемых тем может потребовать десятков и даже сотен запусков.
  Открытая проблема: построение полного набора хорошо интерпретируемых тем за один запуск. Возможно ли это с помощью регуляризации?

- *Моделирование тематики последовательного текста* с учётом порядка слов. Модели, основанные на гипотезе «мешка слов», слишком хаотично относят близко стоящие слова к разным темам. Понятие темы (topic), основанное на статистике сочетаемости слов, плохо согласуется с признанными в лингвистике концепциями темы (theme).
  Открытая проблема: создание тематических моделей внимания, возможно, на основе локализованного E-шага. Методологическая интеграция с лингвистическими методами анализа текста.

- *Развитие нейросетевых тематических моделей* (neural topic model, NTM). Технологическая несовместимость тематических и нейросетевых векторных представлений слов приводит к конкуренции этих концепций, слишком громоздким решениям или слишком поверхностной интеграции.
  Открытая проблема: глубокая интеграция механизмов обучения тематических

и нейросетевых моделей языка с помощью мультипликативных градиентных шагов (основной леммы о максимизации на единичных симплексах).

- *Автоматизация подбора гиперпараметров.* При использовании нескольких регуляризаторов и модальностей их весовые коэффициенты обычно подбираются вручную, путём многократного перестроения модели по всем данным. Применение генетических алгоритмов и других известных техник AutoML проводит к громоздким и вычислительно трудоёмким алгоритмам.
  Открытая проблема: реализация быстрой адаптивной многокритериальной оптимизации гиперпараметров в режиме пакетной обработки коллекции.

- *Автоматическое именование и изложение тем.* Большинство приложений тематического моделирования связано с визуальным представлением тем пользователю. Каждая тема должна уметь рассказать о себе. Однако в литературе задача суммаризации тем до сих пор не ставилась.
  Открытая проблема: генерировать для любой темы не только краткое название и релевантные фразы, но и краткое связное изложение заданного объёма.

- *Динамическое создание событийных тем* в текстовых потоках. Несмотря на обилие исследований по темпоральным тематическим моделям, пока не существует универсальной модели, подходящей для любых текстовых коллекций с динамикой.
  Открытая проблема: автоматическое и своевременное детектирование новой тематической лексики, присутствия новой темы в документе, моментов появления новых тем и завершения старых. Решение этих задач должно обходиться без трудоёмкого экспериментального подбора гиперпараметров.

- *Бережное объединение тематических моделей* при пополнении коллекции значительным объёмом текстовых данных с неизвестным числом новых тем, либо при объединении коллекций с готовыми тематическими моделями.
  Открытая проблема: выделение общих и сохранение уникальных тем при объединении нескольких тематических моделей, возможно, в условиях несбалансированности тем по их объёмам.

О том, что осталось за бортом. В эту книгу не вошли некоторые важные типы моделей, например, для анализа изображений и видеопотоков, аннотирования изображений, суммаризации текстов, анализа тональности и выявления мнений, использования и построения онтологий, обнаружения новых тем и прослеживания новостных сюжетов. Не были затронуты методы автоматического именования тем. Не нашла отражения в обзоре новая тенденция создания гибридных моделей на основе тематического моделирования и нейронных сетей.

Благодарности. Работа выполнена при финансовой поддержке правительства Российской Федерации (соглашение 05.Y09.21.0018), Российского фонда фундаментальных исследований (проекты 17-07-01536, 20-07-00936), Института перспективных исследований проблем искусственного интеллекта и интеллектуальных систем Московского государственного университета имени М. В. Ломоносова.

Автор признателен своим ученикам и коллегам, участвовавшим в разработке библиотеки `BigARTM`, в прикладных проектах с её применением и предоставившим материалы своих экспериментов для этой книги: Василию Алексееву, Мурату Апишеву, Виктору Булатову, Николаю Герасименко, Никите Дойкову, Илье Ирхину, Артёму Попову, Анне Потапенко, Полине Потаповой, Юлиану Сердюку, Михаилу Солоткому, Сергею Стенину, Марине Суворовой, Дмитрию Федоряка, Даниилу Фельдману, Галине Фоминской, Александру Фрею, Кириллу Хрыльченко, Полине Черниковой, Никите Шаповалову, Анастасии Яниной.

Список литературы

[1] *Агеев М. С., Добров Б. В., Лукашевич Н. В.* Автоматическая рубрикация текстов: методы и проблемы // *Учёные записки Казанского государственного университета. Серия Физико-математические науки.* — 2008. — Т. 150, № 4. — С. 25–40.

[2] *Айсина Р. М.* Обзор средств визуализации тематических моделей коллекций текстовых документов // *Машинное обучение и анализ данных (http://jmlda.org).* — 2015. — Т. 1, № 11. — С. 1584–1618.

[3] *Апишев М. А.* Эффективные реализации алгоритмов тематического моделирования // *Труды ИСП РАН.* — 2020. — Т. 32, № 1. — С. 137–152.

[4] *ван Дейк Т.* Язык. Познание. Коммуникация. Пер. с англ. Изд. 2. — Благовещенск: Благовещенский гуманитарный колледж, 2000. — 310 с.

[5] *Воронцов К. В.* Аддитивная регуляризация тематических моделей коллекций текстовых документов // *Доклады РАН.* — 2014. — Т. 456, № 3. — С. 268–271.

[6] *Воронцов К. В., Потапенко А. А.* Регуляризация, робастность и разреженность вероятностных тематических моделей // *Компьютерные исследования и моделирование.* — 2012. — Т. 4, № 4. — С. 693–706.

[7] *Воронцов К. В., Потапенко А. А.* Модификации EM-алгоритма для вероятностного тематического моделирования // *Машинное обучение и анализ данных.* — 2013. — Т. 1, № 6. — С. 657–686.

[8] *Воронцов К. В., Потапенко А. А.* Регуляризация вероятностных тематических моделей для повышения интерпретируемости и определения числа тем // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 4–8 июня 2014 г.). — Вып. 13 (20). — М: Изд-во РГГУ, 2014. — С. 676–687.

[9] *Воронцов К. В., Фрей А. И., Ромов П. А., Янина А. О., Суворова М. А., Апишев М. А.* BigARTM: библиотека с открытым кодом для тематического моделирования больших текстовых коллекций // Аналитика и управление данными в областях с интенсивным использованием данных. XVII Международная конференция DAMDID/RCDL'2015. — НИЯУ МИФИ Обнинск, 2015. — С. 28–36.

[10] *Дударенко М. А.* Регуляризация многоязычных тематических моделей // *Вычислительные методы и программирование.* — 2015. — Т. 16. — С. 26–38.

[11] *Ирхин И. А., Булатов В. Г., Воронцов К. В.* Аддитивная регуляризация тематических моделей с быстрой векторизацией текста // *Компьютерные исследования и моделирование.* — 2020. — Т. 12, № 6. — С. 1515–1528.

[12] *Ирхин И. А., Воронцов К. В.* Сходимость алгоритма аддитивной регуляризации тематических моделей // *Труды Института математики и механики УрО РАН.* — 2020. — Т. 26, № 3. — С. 56–68.

[13] *Колмогоров А. Н.* Теория информации и теория алгоритмов / Под ред. Ю. В. Прохоров. — М.: Наука, 1987. — 304 с.

[14] *Лукашевич Н. В.* Тезаурусы в задачах информационного поиска. — Издательство МГУ имени М. В. Ломоносова, 2011.

[15] *Маннинг К. Д., Рагхаван П., Шютце Х.* Введение в информационный поиск. — Вильямс, 2011.

[16] *Павлов А. С., Добров Б. В.* Метод обнаружения массово порожденных неестественных текстов на основе анализа тематической структуры // *Вычислительные методы и программирование: новые вычислительные технологии.* — 2011. — Т. 12. — С. 58–72.

[17] *Тихонов А. Н., Арсенин В. Я.* Методы решения некорректных задач. — М.: Наука, 1986.

[18] *Хрыльченко К. Я., Воронцов К. В.* Оптимизация весов модальностей в тематических моделях транзакционных данных // *Автоматика и телемеханика.* — 2022. — № 12. — С. 44–62.

[19] *Целых В. Р., Воронцов К. В.* Критерии согласия для разреженных дискретных распределений и их применение в тематическом моделировании // *Машинное обучение и анализ данных.* — 2012. — Т. 1, № 4. — С. 437–447.

[20] *Янина А. О., Воронцов К. В.* Мультимодальные тематические модели для разведочного поиска в коллективном блоге // *Машинное обучение и анализ данных.* — 2016. — Т. 2, № 2. — С. 173–186.

[21] *Airoldi E. M., Erosheva E. A., Fienberg S. E., Joutard C., Love T., Shringarpure S.* Reconceptualizing the classification of PNAS articles // *Proceedings of The National Academy of Sciences.* — 2010. — Vol. 107. — Pp. 20899–20904.

[22] *Altaweel M., Bone C., Abrams J.* Documents as data: A content analysis and topic modeling approach for analyzing responses to ecological disturbances // *Ecological Informatics.* — 2019. — Vol. 51. — Pp. 82–95.

[23] *Andrzejewski D., Buttler D.* Latent topic feedback for information retrieval // Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. — KDD '11. — 2011. — Pp. 600–608.

[24] *Andrzejewski D., Zhu X.* Latent Dirichlet allocation with topic-in-set knowledge // Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing. — SemiSupLearn '09. — Stroudsburg, PA, USA: Association for Computational Linguistics, 2009. — Pp. 43–48.

[25] *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.

[26] *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.

[27] *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.

[28] *Asuncion A., Welling M., Smyth P., Teh Y. W.* On smoothing and inference for topic models // Proceedings of the International Conference on Uncertainty in Artificial Intelligence. — 2009. — Pp. 27–34.

[29] 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.

[30] *Bakharia A.* On the equivalence of inductive content analysis and topic modeling // Advances in Quantitative Ethnography / Ed. by B. Eagan, M. Misfeldt, A. Siebert-Evenstone. — Cham: Springer International Publishing, 2019. — Pp. 291–298.

[31] *Bakharia A., Bruza P., Watters J., Narayan B., Sitbon L.* Interactive topic modeling for aiding qualitative content analysis // Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval. — CHIIR '16. — New York, NY, USA: Association for Computing Machinery, 2016. — P. 213–222.

[32] *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.

[33] *Bassiou N., Kotropoulos C.* Online PLSA: Batch updating techniques including out-of-vocabulary words // *Neural Networks and Learning Systems, IEEE Transactions on.* — Nov 2014. — Vol. 25, no. 11. — Pp. 1953–1966.

[34] *Bishop C. M.* Pattern Recognition and Machine Learning. — Springer, Series: Information Science and Statistics, 2006. — 740 pp.

[35] *Blei D., Lafferty J.* A correlated topic model of Science // *Annals of Applied Statistics.* — 2007. — Vol. 1. — Pp. 17–35.

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

[37] *Blei D. M., Griffiths T., Jordan M., Tenenbaum J.* Hierarchical topic models and the nested chinese restaurant process // NIPS. — 2003.

[38] *Blei D. M., Griffiths T. L., Jordan M. I.* The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies // *J. ACM.* — 2010. — Vol. 57, no. 2. — Pp. 7:1–7:30.

[39] *Blei D. M., Jordan M. I.* Modeling annotated data // Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval. — New York, NY, USA: ACM, 2003. — Pp. 127–134.

[40] *Blei D. M., Ng A. Y., Jordan M. I.* Latent dirichlet allocation // Advances in Neural Information Processing Systems 14, NIPS 2001, December 3–8, 2001, Vancouver, British Columbia, Canada] / Ed. by T. G. Dietterich, S. Becker, Z. Ghahramani. — MIT Press, 2001. — Pp. 601–608.

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

[42] *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.

[43] *Bolelli L., Ertekin S., Giles C. L.* Topic and trend detection in text collections using latent Dirichlet allocation // ECIR. — Vol. 5478 of *Lecture Notes in Computer Science.* — Springer, 2009. — Pp. 776–780.

[44] *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.

[45] *Chang J., Gerrish S., Wang C., Boyd-Graber J. L., Blei D. M.* Reading tea leaves: How humans interpret topic models // Neural Information Processing Systems (NIPS). — 2009. — Pp. 288–296.

[46] *Chemudugunta C., Smyth P., Steyvers M.* Modeling general and specific aspects of documents with a probabilistic topic model // Advances in Neural Information Processing Systems. — Vol. 19. — MIT Press, 2007. — Pp. 241–248.

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

[48] *Chien J.-T., Chang Y.-L.* Bayesian sparse topic model // *Journal of Signal Processessing Systems.* — 2013. — Vol. 74. — Pp. 375–389.

[49] *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.

[50] *Chuang J., Gupta S., Manning C., Heer J.* Topic model diagnostics: Assessing domain relevance via topical alignment // Proceedings of the 30th International Conference on Machine Learning (ICML-13) / Ed. by S. Dasgupta, D. Mcallester. — Vol. 28. — JMLR Workshop and Conference Proceedings, 2013. — Pp. 612–620.

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

[52] Computer-assisted content analysis: Topic models for exploring multiple subjective interpretations // NIPS Workshop on Human-Propelled Machine Learning. — 2014.

[53] *Cressie N., Read T. R. C.* Multinomial goodness-of-fit tests // *Journal of the Royal Statistical Society, Series B.* — 1984. — Vol. 46, no. 3. — Pp. 440–464.

[54] *Dai A. M., Olah C., Le Q. V.* Document embedding with paragraph vectors // NIPS Deep Learning Workshop. — 2015.

[55] *Daud A., Li J., Zhou L., Muhammad F.* Knowledge discovery through directed probabilistic topic models: a survey // *Frontiers of Computer Science in China.* — 2010. — Vol. 4, no. 2. — Pp. 280–301.

[56] *De Smet W., Moens M.-F.* Cross-language linking of news stories on the web using interlingual topic modelling // Proceedings of the 2Nd ACM Workshop on Social Web Search and Mining. — SWSM '09. — New York, NY, USA: ACM, 2009. — Pp. 57–64.

[57] *Dempster A. P., Laird N. M., Rubin D. B.* Maximum likelihood from incomplete data via the EM algorithm // *J. of the Royal Statistical Society, Series B.* — 1977. — no. 34. — Pp. 1–38.

[58] *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.

[59] *Dietz L., Bickel S., Scheffer T.* Unsupervised prediction of citation influences // Proceedings of the 24th international conference on Machine learning. — ICML '07. — New York, NY, USA: ACM, 2007. — Pp. 233–240.

[60] *Doyle G., Elkan C.* Accounting for burstiness in topic models // Proceedings of the 26th Annual International Conference on Machine Learning. — ICML'09. — New York, NY, USA: ACM, 2009. — Pp. 281–288.

[61] *Egghe L.* Untangling Herdan's law and Heaps' law: Mathematical and informetric arguments // *Journal of the American Society for Information Science and Technology.* — 2007. — Vol. 58, no. 5. — Pp. 702–709.

[62] *Eisenstein J., Ahmed A., Xing E. P.* Sparse additive generative models of text // ICML'11. — 2011. — Pp. 1041–1048.

[63] *Ejaz W., Ittefaq M., Jamil S.* Politics triumphs: a topic modeling approach of analyzing news media coverage of climate change in pakistan // *Journal of Science Communication.* — 01 2023. — Vol. 22. — P. A02.

[64] *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.

[65] *Fan A., Doshi-Velez F., Miratrix L.* Assessing topic model relevance: Evaluation and informative priors // *Statistical Analysis and Data Mining: The ASA Data Science Journal.* — 2019. — Vol. 12, no. 3. — Pp. 210–222.

[66] *Feldman S. E.* The answer machine // Synthesis Lectures on Information Concepts, Retrieval, and Services. — Morgan & Claypool Publishers, 2012. — Vol. 4. — Pp. 1–137.

[67] *Feng Y., Lapata M.* Topic models for image annotation and text illustration // Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. — Association for Computational Linguistics, 2010. — Pp. 831–839.

[68] *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.

[69] *Girolami M., Kabán A.* On an equivalence between PLSI and LDA // SIGIR'03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval. — 2003. — Pp. 433–434.

[70] *Goldstone A., Underwood T.* The quiet transformations of literary studies: What thirteen thousand scholars could tell us // *New Literary History.* — 2014. — Vol. 45, no. 3. — Pp. 359–384.

[71] *Grant C. E., George C. P., Kanjilal V., Nirkhiwale S., Wilson J. N., Wang D. Z.* A topic-based search, visualization, and exploration system // FLAIRS Conference. — AAAI Press, 2015. — Pp. 43–48.

[72] *Hagen L.* Content analysis of e-petitions with topic modeling: How to train and evaluate LDA models? // *Inf. Process. Manag.* — 2018. — Vol. 54, no. 6. — Pp. 1292–1307.

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

[74] *He Z., Mokhberian N., Camara A., Abeliuk A., Lerman K.* Detecting polarized topics using partisanship-aware contextualized topic embeddings // Conference on Empirical Methods in Natural Language Processing. — Association for Computational Linguistics, 2021. — Pp. 2102–2118.

[75] *Helsgaun K.* An effective implementation of the linвЂ“kernighan traveling salesman heuristic // *European Journal of Operational Research.* — 2000. — Vol. 126, no. 1. — Pp. 106–130.

[76] *Hoffman M. D., Blei D. M., Bach F. R.* Online learning for latent Dirichlet allocation // NIPS. — Curran Associates, Inc., 2010. — Pp. 856–864.

[77] *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.

[78] *Hospedales T., Gong S., Xiang T.* Video behaviour mining using a dynamic topic model // *International Journal of Computer Vision.* — 2012. — Vol. 98, no. 3. — Pp. 303–323.

[79] *Huang P.-S., He X., Gao J., Deng L., Acero A., Heck L.* Learning deep structured semantic models for web search using clickthrough data // Proceedings of the 22Nd ACM International Conference on Conference on Information and Knowledge Management. — CIKM '13. — New York, NY, USA: ACM, 2013. — Pp. 2333–2338.

[80] *Ianina 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.

[81] *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.

[82] *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.

[83] *Jacksi K., Dimililer N., Zeebaree S. R. M.* A survey of exploratory search systems based on LOD resources // Proceedings of the 5th International Conference on Computing and Informatics, ICOCI 2015. — School of Computing, Universiti Utara Malaysia, 2015. — Pp. 501–509.

[84] *Jagarlamudi J., Daumé III H., Udupa R.* Incorporating lexical priors into topic models // Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics. — EACL'12. — Stroudsburg, PA, USA: Association for Computational Linguistics, 2012. — Pp. 204–213.

[85] *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.

[86] *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.

[87] *Jiang T.* Exploratory Search: A Critical Analysis of the Theoretical Foundations, System Features, and Research Trends // Library and Information Sciences: Trends and Research / Ed. by C. Chen, R. Larsen. — Berlin, Heidelberg: Springer Berlin Heidelberg, 2014. — Pp. 79–103.

[88] *Kataria S., Mitra P., Caragea C., Giles C. L.* Context sensitive topic models for author influence in document networks // Proceedings of the Twenty-Second international joint conference on Artificial Intelligence — Volume 3. — IJCAI'11. — AAAI Press, 2011. — Pp. 2274–2280.

[89] *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.

[90] *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.

[91] *Kim S.-J., Koh K., Boyd S., Gorinevsky D.* L1 trend filtering // *SIAM review*. — 2009. — Vol. 51, no. 2. — Pp. 339–360.

[92] *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.

[93] *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.

[94] *Konietzny S., Dietz L., McHardy A.* Inferring functional modules of protein families with probabilistic topic models // *BMC Bioinformatics*. — 2011. — Vol. 12, no. 1. — P. 141.

[95] *Krestel R., Fankhauser P., Nejdl W.* Latent Dirichlet allocation for tag recommendation // Proceedings of the third ACM conference on Recommender systems. — ACM, 2009. — Pp. 61–68.

[96] *La Rosa M., Fiannaca A., Rizzo R., Urso A.* Probabilistic topic modeling for the analysis and classification of genomic sequences // *BMC Bioinformatics*. — 2015. — Vol. 16, no. Suppl 6. — P. S2.

[97] *Lample G., Ballesteros M., Subramanian S., Kawakami K., Dyer C.* Neural architectures for named entity recognition // HLT-NAACL / Ed. by K. Knight, A. Nenkova, O. Rambow. — The Association for Computational Linguistics, 2016. — Pp. 260–270.

[98] 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.

[99] *Larsson M. O., Ugander J.* A concave regularization technique for sparse mixture models // Advances in Neural Information Processing Systems 24 / Ed. by J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, K. Weinberger. — 2011. — Pp. 1890–1898.

[100] *Lee S. S., Chung T., McLeod D.* Dynamic item recommendation by topic modeling for social networks // Information Technology: New Generations (ITNG), 2011 Eighth International Conference on. — IEEE, 2011. — Pp. 884–889.

[101] *Lei S., Zhang J., Weng S., Zhang C.* Topic model with constrainted word burstiness intensities // The 2011 International Joint Conference on Neural Networks. — 2011. — Pp. 68–74.

[102] *Levy O., Goldberg Y.* Neural word embedding as implicit matrix factorization // Advances in Neural Information Processing Systems 27 / Ed. by Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, K. Q. Weinberger. — Curran Associates, Inc., 2014. — Pp. 2177–2185.

[103] *Li S., Li J., Pan R.* Tag-weighted topic model for mining semi-structured documents // IJCAI'13 Proceedings of the Twenty-Third international joint conference on Artificial Intelligence. — AAAI Press, 2013. — Pp. 2855–2861.

[104] *Li W., McCallum A.* Pachinko allocation: Dag-structured mixture models of topic correlations // ICML. — 2006.

[105] *Li X.-X., Sun C.-B., Lu P., Wang X.-J., Zhong Y.-X.* Simultaneous image classification and annotation based on probabilistic model // *The Journal of China Universities of Posts and Telecommunications.* — 2012. — Vol. 19, no. 2. — Pp. 107–115.

[106] *Litvak M., Vanetik N., Liu C., Xiao L., Savas O.* Improving summarization quality with topic modeling // Proceedings of the 2015 Workshop on Topic Models: Post-Processing and Applications. — New York, NY, USA: Association for Computing Machinery, 2015. — Pp. 39–47.

[107] *Liu J., Shang J., Wang C., Ren X., Han J.* Mining quality phrases from massive text corpora // Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data. — SIGMOD '15. — New York, NY, USA: ACM, 2015. — Pp. 1729–1744.

[108] *Liu Y., Liu Z., Chua T.-S., Sun M.* Topical word embeddings // Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence. — AAAI'15. — AAAI Press, 2015. — Pp. 2418–2424.

[109] *Lu Y., Mei Q., Zhai C.* Investigating task performance of probabilistic topic models: an empirical study of PLSA and LDA // *Information Retrieval.* — 2011. — Vol. 14, no. 2. — Pp. 178–203.

[110] *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.

[111] *Mann G. S., McCallum A.* Simple, robust, scalable semi-supervised learning via expectation regularization // Proceedings of the 24th international conference on Machine learning. — ICML '07. — New York, NY, USA: ACM, 2007. — Pp. 593–600.

[112] *Marchionini G.* Exploratory search: From finding to understanding // *Commun. ACM.* — 2006. — Vol. 49, no. 4. — Pp. 41–46.

[113] *Marie N., Gandon F.* Survey of linked data based exploration systems // Proceedings of the 3rd International Workshop on Intelligent Exploration of Semantic Data (IESD 2014) co-located with the 13th International Semantic Web Conference (ISWC 2014), Riva del Garda, Italy, October 20, 2014. — 2014.

[114] *Marjanen J., Zosa E., Hengchen S., Pivovarova L., Tolonen M.* Topic modelling discourse dynamics in historical newspapers // *CoRR.* — 2020. — Vol. abs/2011.10428.

[115] *Masada T., Kiyasu S., Miyahara S.* Comparing LDA with pLSI as a dimensionality reduction method in document clustering // Proceedings of the 3rd International Conference on Large-scale knowledge resources: construction and application. — LKR'08. — Springer-Verlag, 2008. — Pp. 13–26.

[116] *McAuliffe J. D., Blei D. M.* Supervised topic models // Advances in Neural Information Processing Systems 20 / Ed. by J. C. Platt, D. Koller, Y. Singer, S. T. Roweis. — Curran Associates, Inc., 2008. — Pp. 121–128.

[117] *Mei Q., Cai D., Zhang D., Zhai C.* Topic modeling with network regularization // Proceedings of the 17th International Conference on World Wide Web. — WWW'08. — New York, NY, USA: ACM, 2008. — Pp. 101–110.

[118] *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.

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

[120] *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.

[121] *Mimno D., Blei D.* Bayesian checking for topic models // 11th Conference on Empirical Methods in Natural Language Processing. — Association for Computational Linguistics, 2011. — Pp. 227–237.

[122] *Mimno D., Hoffman M., Blei D.* Sparse stochastic inference for latent Dirichlet allocation // Proceedings of the 29th International Conference on Machine Learning (ICML-12) / Ed. by J. Langford, J. Pineau. — New York, NY, USA: Omnipress, July 2012. — Pp. 1599–1606.

[123] *Mimno D., Li W., McCallum A.* Mixtures of hierarchical topics with pachinko allocation // ICML. — 2007.

[124] *Mimno D., Wallach H. M., Naradowsky J., Smith D. A., McCallum A.* Polylingual topic models // Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2 - Volume 2. — EMNLP '09. — Stroudsburg, PA, USA: Association for Computational Linguistics, 2009. — Pp. 880–889.

[125] *Mimno D., Wallach H. M., Talley E., Leenders M., McCallum A.* Optimizing semantic coherence in topic models // Proceedings of the Conference on Empirical Methods in Natural Language Processing. — EMNLP '11. — Stroudsburg, PA, USA: Association for Computational Linguistics, 2011. — Pp. 262–272.

[126] *Mimno D. M.* Computational historiography: Data mining in a century of classics journals. // *ACM Journal on Computing and Cultural Heritage*. — 2012. — Vol. 5, no. 1. — Pp. 3:1–3:19.

[127] *Minka T. P.* Estimating a Dirichlet distribution: Tech. rep.: 2000 (revised 2003, 2009, 2012).

[128] *Moretti F.* Graphs, maps, trees : abstract models for literary history. — London; New York: Verso, 2007.

[129] *Nadeau D., Sekine S.* A survey of named entity recognition and classification // *Linguisticae Investigationes*. — 2007. — Vol. 30, no. 1. — Pp. 3–26.

[130] *Newman D., Bonilla E. V., Buntine W. L.* Improving topic coherence with regularized topic models // Advances in Neural Information Processing Systems 24 / Ed. by J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, K. Weinberger. — 2011. — Pp. 496–504.

[131] *Newman D., Chemudugunta C., Smyth P.* Statistical entity-topic models // Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. — KDD '06. — New York, NY, USA: ACM, 2006. — Pp. 680–686.

[132] *Newman D., Karimi S., Cavedon L.* External evaluation of topic models // Australasian Document Computing Symposium. — December 2009. — Pp. 11–18.

[133] *Newman D., Lau J. H., Grieser K., Baldwin T.* Automatic evaluation of topic coherence // Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. — HLT '10. — Stroudsburg, PA, USA: Association for Computational Linguistics, 2010. — Pp. 100–108.

[134] *Newman D., Noh Y., Talley E., Karimi S., Baldwin T.* Evaluating topic models for digital libraries // Proceedings of the 10th annual Joint Conference on Digital libraries. — JCDL '10. — New York, NY, USA: ACM, 2010. — Pp. 215–224.

[135] *Ni J., Dinu G., Florian R.* Weakly supervised cross-lingual named entity recognition via effective annotation and representation projection // The 55th Annual Meeting of the Association for Computational Linguistics (ACL). — 2017.

[136] *Ni X., Sun J.-T., Hu J., Chen Z.* Mining multilingual topics from wikipedia // Proceedings of the 18th International Conference on World Wide Web. — WWW '09. — New York, NY, USA: ACM, 2009. — Pp. 1155–1156.

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

[138] 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.

[139] *Palagi E., Gandon F., Giboin A., Troncy R.* A survey of definitions and models of exploratory search // ESIDA'17 - ACM Workshop on Exploratory Search and Interactive Data Analytics, Mar 2017, Limassol, Cyprus. — 03 2017. — Pp. 3–8.

[140] *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.

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

[142] *Paul M. J., Girju R.* Topic modeling of research fields: An interdisciplinary perspective // RANLP. — RANLP 2009 Organising Committee / ACL, 2009. — Pp. 337–342.

[143] *Pennington J., Socher R., Manning C. D.* GloVe: Global vectors for word representation // Empirical Methods in Natural Language Processing (EMNLP). — 2014. — Pp. 1532–1543.

[144] *Phuong D. V., Phuong T. M.* A keyword-topic model for contextual advertising // Proceedings of the Third Symposium on Information and Communication Technology. — SoICT '12. — New York, NY, USA: ACM, 2012. — Pp. 63–70.

[145] *Pinto J. C. L., Chahed T.* Modeling multi-topic information diffusion in social networks using latent Dirichlet allocation and Hawkes processes // Tenth International Conference on Signal-Image Technology & Internet-Based Systems. — 2014. — Pp. 339–346.

[146] *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.

[147] *Potapenko A. A., Vorontsov K. V.* Robust PLSA performs better than LDA // 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. 784–787.

[148] *Pritchard J. K., Stephens M., Donnelly P.* Inference of population structure using multilocus genotype data // *Genetics*. — 2000. — Vol. 155. — Pp. 945–959.

[149] *Pujara J., Skomoroch P.* Large-scale hierarchical topic models // NIPS Workshop on Big Learning. — 2012.

[150] *Rahman M.* Search engines going beyond keyword search: A survey // *International Journal of Computer Applications*. — August 2013. — Vol. 75, no. 17. — Pp. 1–8.

[151] *Ramage D., Hall D., Nallapati R., Manning C. D.* Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora // Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1. — EMNLP '09. — Stroudsburg, PA, USA: Association for Computational Linguistics, 2009. — Pp. 248–256.

[152] *Reisenbichler M., Reutterer T.* Topic modeling in marketing: recent advances and research opportunities // *Journal of Business Economics*. — 2019. — Vol. 89, no. 3. — Pp. 327–356.

[153] *Riedl M., Biemann C.* TopicTiling: A text segmentation algorithm based on LDA // Proceedings of ACL 2012 Student Research Workshop. — ACL '12. — Stroudsburg, PA, USA: Association for Computational Linguistics, 2012. — Pp. 37–42.

[154] *Rönnqvist S.* Exploratory topic modeling with distributional semantics // Advances in Intelligent Data Analysis XIV: 14th International Symposium, IDA 2015, Saint Etienne. France, October 22 -24, 2015. Proceedings / Ed. by E. Fromont, T. De Bie, M. van Leeuwen. — Springer International Publishing, 2015. — Pp. 241–252.

[155] *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.

[156] *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.

[157] *Scherer M., von Landesberger T., Schreck T.* Topic modeling for search and exploration in multivariate research data repositories // Research and Advanced Technology for Digital Libraries: International Conference on Theory and Practice of Digital Libraries, TPDL 2013, Valletta, Malta, September 22-26, 2013. Proceedings / Ed. by T. Aalberg, C. Papatheodorou, M. Dobreva, G. Tsakonas, C. J. Farrugia. — Berlin, Heidelberg: Springer Berlin Heidelberg, 2013. — Pp. 370–373.

[158] *Schönfeld M., Eckhard S., Patz R., van Meegdenburg H.* Discursive landscapes and unsupervised topic modeling in ir: A validation of text-as-data approaches through a new corpus of un security council speeches on afghanistan // *ArXiv*. — 2018. — Vol. abs/1810.05572.

[159] *Shadrova A.* Topic models do not model topics: epistemological remarks and steps towards best practices // *Journal of Data Mining & Digital Humanities*. — 2021. — Vol. 2021.

[160] *Shang J., Liu J., Jiang M., Ren X., Voss C. R., Han J.* Automated phrase mining from massive text corpora // *CoRR*. — 2017. — Vol. abs/1702.04457.

[161] *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.

[162] *Shashanka M., Raj B., Smaragdis P.* Sparse overcomplete latent variable decomposition of counts data // Advances in Neural Information Processing Systems, NIPS-2007 / Ed. by J. C. Platt, D. Koller, Y. Singer, S. Roweis. — Cambridge, MA: MIT Press, 2008. — Pp. 1313–1320.

[163] *Shivashankar S., Srivathsan S., Ravindran B., Tendulkar A. V.* Multi-view methods for protein structure comparison using latent dirichlet allocation. // *Bioinformatics [ISMB/ECCB]*. — 2011. — Vol. 27, no. 13. — Pp. 61–68.

[164] *Shneiderman B.* The eyes have it: A task by data type taxonomy for information visualizations // Proceedings of the 1996 IEEE Symposium on Visual Languages. — VL'96. — Washington, DC, USA: IEEE Computer Society, 1996. — Pp. 336–343.

[165] *Si X., Sun M.* Tag-LDA for scalable real-time tag recommendation // *Journal of Information & Computational Science*. — 2009. — Vol. 6. — Pp. 23–31.

[166] *Singh R., Hsu Y.-W., Moon N.* Multiple perspective interactive search: a paradigm for exploratory search and information retrieval on the Web // *Multimedia Tools and Applications*. — 2013. — Vol. 62, no. 2. — Pp. 507–543.

[167] *Sokolov E., Bogolubsky L.* Topic models regularization and initialization for regression problems // Proceedings of the 2015 Workshop on Topic Models: Post-Processing and Applications. — New York, NY, USA: ACM, 2015. — Pp. 21–27.

[168] *Steyvers M., Griffiths T.* Finding scientific topics // *Proceedings of the National Academy of Sciences*. — 2004. — Vol. 101, no. Suppl. 1. — Pp. 5228–5235.

[169] *Sun Y., Han J., Gao J., Yu Y.* iTopicModel: Information network-integrated topic modeling // 2009 Ninth IEEE International Conference on Data Mining. — 2009. — Pp. 493–502.

[170] *Taddy M.* On estimation and selection for topic models // Artificial Intelligence and Statistics. — 2012. — Pp. 1184–1193.

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

[172] *Teh Y. W., Jordan M. I., Beal M. J., Blei D. M.* Hierarchical Dirichlet processes // *Journal of the American Statistical Association*. — 2006. — Vol. 101, no. 476. — Pp. 1566–1581.

[173] *Teh Y. W., Newman D., Welling M.* A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation // NIPS. — 2006. — Pp. 1353–1360.

[174] TextFlow: Towards better understanding of evolving topics in text. / W. Cui, S. Liu, L. Tan, C. Shi, Y. Song, Z. Gao, H. Qu, X. Tong // *IEEE transactions on visualization and computer graphics*. — 2011. — Vol. 17, no. 12. — Pp. 2412–2421.

[175] *Varadarajan J., Emonet R., Odobez J.-M.* A sparsity constraint for topic models — application to temporal activity mining // NIPS-2010 Workshop on Practical Applications of Sparse Modeling: Open Issues and New Directions. — 2010.

[176] *Varshney D., Kumar S., Gupta V.* Modeling information diffusion in social networks using latent topic information // Intelligent Computing Theory / Ed. by D.-S. Huang, V. Bevilacqua, P. Premaratne. — Springer International Publishing, 2014. — Vol. 8588 of *Lecture Notes in Computer Science*. — Pp. 137–148.

[177] *Veas E. E., di Sciascio C.* Interactive topic analysis with visual analytics and recommender systems // 2nd Workshop on Cognitive Computing and Applications for Augmented Human Intelligence, CCAAHI2015, International Joint Conference on Artificial Intelligence, IJCAI, Buenos Aires, Argentina, July 2015. — Aachen, Germany, Germany: CEUR-WS.org, 2015.

[178] *Vorontsov K., Frei O., Apishev M., Romov P., Suvorova M., Ianina 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.

[179] *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.

[180] *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.

[181] *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.

[182] *Vorontsov K. V., Potapenko A. A., Plavin A. V.* Additive regularization of topic models for topic selection and sparse factorization // The Third International Symposium On Learning And Data Sciences (SLDS 2015). April 20-22, 2015. Royal Holloway, University of London, UK. / Ed. by A. G. et al. — Springer International Publishing Switzerland 2015, 2015. — Pp. 193–202.

[183] *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.

[184] *Vulić I., Smet W., Moens M.-F.* Cross-language information retrieval models based on latent topic models trained with document-aligned comparable corpora // *Information Retrieval*. — 2012. — Pp. 1–38.

[185] *Wallach H.* Structured Topic Models for Language: Ph.D. thesis / Newnham College, University of Cambridge. — 2008.

[186] *Wallach H., Mimno D., McCallum A.* Rethinking LDA: Why priors matter // Advances in Neural Information Processing Systems 22: 23rd Annual Conference on Neural Information Processing Systems, Vancouver, BC, Canada / Ed. by Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, A. Culotta. — 2009. — Pp. 1973–1981.

[187] *Wallach H., Murray I., Salakhutdinov R., Mimno D.* Evaluation methods for topic models // 26th International Conference on Machine Learning, Montreal, Canada. — 2009. — Pp. 1105–1112.

[188] *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.

[189] *Wang C., Blei D. M.* Decoupling sparsity and smoothness in the discrete hierarchical Dirichlet process // NIPS. — Curran Associates, Inc., 2009. — Pp. 1982–1989.

[190] *Wang C., Blei D. M.* Collaborative topic modeling for recommending scientific articles // Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. — New York, NY, USA: ACM, 2011. — Pp. 448–456.

[191] *Wang C., Danilevsky M., Desai N., Zhang Y., Nguyen P., Taula T., Han J.* A phrase mining framework for recursive construction of a topical hierarchy // Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. — KDD '13. — New York, NY, USA: ACM, 2013. — Pp. 437–445.

[192] *Wang C., Liu J., Desai N., Danilevsky M., Han J.* Constructing topical hierarchies in heterogeneous information networks // *Knowledge and Information Systems.* — 2014. — Vol. 44, no. 3. — Pp. 529–558.

[193] *Wang C., Liu X., Song Y., Han J.* Scalable and robust construction of topical hierarchies // *CoRR.* — 2014. — Vol. abs/1403.3460.

[194] *Wang C., Liu X., Song Y., Han J.* Towards interactive construction of topical hierarchy: A recursive tensor decomposition approach // Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. — KDD '15. — New York, NY, USA: ACM, 2015. — Pp. 1225–1234.

[195] *Wang H., Zhang D., Zhai C.* Structural topic model for latent topical structure analysis // Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1. — HLT '11. — Stroudsburg, PA, USA: Association for Computational Linguistics, 2011. — Pp. 1526–1535.

[196] *Wang X., McCallum A.* Topics over time: A non-markov continuous-time model of topical trends // Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. — KDD '06. — New York, NY, USA: ACM, 2006. — Pp. 424–433.

[197] *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.

[198] *Wesslen R.* Computer-assisted text analysis for social science: Topic models and beyond // *CoRR.* — 2018. — Vol. abs/1803.11045.

[199] *White R. W., Roth R. A.* Exploratory Search: Beyond the Query-Response Paradigm. Synthesis Lectures on Information Concepts, Retrieval, and Services. — Morgan and Claypool Publishers, 2009.

[200] *Wu Y., Ding Y., Wang X., Xu J.* A comparative study of topic models for topic clustering of Chinese web news // Computer Science and Information Technology (ICCSIT), 2010 3rd IEEE International Conference on. — Vol. 5. — july 2010. — Pp. 236–240.

[201] *Yan X., Guo J., Lan Y., Cheng X.* A biterm topic model for short texts // Proceedings of the 22Nd International Conference on World Wide Web. — WWW '13. — Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2013. — Pp. 1445–1456.

[202] *Yeh J.-h., Wu M.-l.* Recommendation based on latent topics and social network analysis // Proceedings of the 2010 Second International Conference on Computer Engineering and Applications. — Vol. 1. — IEEE Computer Society, 2010. — Pp. 209–213.

[203] *Yi X., Allan J.* A comparative study of utilizing topic models for information retrieval // Advances in Information Retrieval. — Springer Berlin Heidelberg, 2009. — Vol. 5478 of *Lecture Notes in Computer Science.* — Pp. 29–41.

[204] *Yin H., Cui B., Chen L., Hu Z., Zhang C.* Modeling location-based user rating profiles for personalized recommendation // *ACM Transactions of Knowledge Discovery from Data.* — 2015.

[205] *Yin H., Cui B., Sun Y., Hu Z., Chen L.* LCARS: A spatial item recommender system // *ACM Transaction on Information Systems.* — 2014.

[206] *Yin Z., Cao L., Han J., Zhai C., Huang T.* Geographical topic discovery and comparison // Proceedings of the 20th international conference on World wide web / ACM. — 2011. — Pp. 247–256.

[207] *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.

[208] *Zhang J., Song Y., Zhang C., Liu S.* Evolutionary hierarchical Dirichlet processes for multiple correlated time-varying corpora // Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. — 2010. — Pp. 1079–1088.

[209] *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.

[210] *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.

[211] *Zhao X. W., Wang J., He Y., Nie J.-Y., Li X.* Originator or propagator?: Incorporating social role theory into topic models for Twitter content analysis // Proceedings of the 22Nd ACM International Conference on Conference on Information and Knowledge Management. — CIKM '13. — New York, NY, USA: ACM, 2013. — Pp. 1649–1654.

[212] *Zhou S., Li K., Liu Y.* Text categorization based on topic model // *International Journal of Computational Intelligence Systems.* — 2009. — Vol. 2, no. 4. — Pp. 398–409.

[213] *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.