Вероятностное тематическое моделирование: теория, модели, алгоритмы и проект BigARTM

Воронцов Константин Вячеславович

vokov@forecsys.ru

Московский физико-технический институт (государственный университет),

Федеральный исследовательский центр «Информатика и управление» РАН

22 февраля 2021 г.

Содержание

1 Введение

Тематическое моделирование — одно из современных направлений *обработки естественного языка* (natural language processing, NLP), активно развивающееся 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.

Вероятностная тематическая модель (probabilistic topic model, PTM) описывает каждую тему дискретным распределением вероятностей слов, а каждый документ — дискретным распределением вероятностей тем. Тематическая модель преобразует любой текст в вектор вероятностей тем. Похожую задачу решают модели *векторных представлений слов* (word embedding) [94, 79], предложений [112, 169] и документов [44]. Особенность вероятностных тематических векторных представлений текста в том, что они интерпретируемые и разреженные.

Тематическое моделирование похоже также на *кластеризацию документов* (document clustering). Отличие в том, что при кластеризации документ целиком относится к одному кластеру, тогда как тематическая модель осуществляет *мягкую кластеризацию* (soft clustering), разделяя документ между несколькими кластерами-темами. Тематические модели называют также моделями мягкой би-кластеризации, поскольку слова также кластеризуются по темам. Это позволяет обходить проблемы синонимии и полисемии слов. Синонимы, употребляемые в схожих контекстах, группируются в одних и тех же темах. Многозначные слова и омонимы, наоборот, распределяют свои вероятности по нескольким семантически не связанным темам.

Многие приложения текстовой аналитики используют тематические векторные представления текста: выявление трендов в новостных потоках, патентных базах, научных публикациях [179, 144], многоязычный информационный поиск [154, 153], классификация и категоризация документов [129, 182], тематическая сегментация текстов [165, 126], суммаризация текстов [83], поиск тематических сообществ в социальных сетях [181, 146, 118, 34], тегирование веб-страниц [73], обнаружение текстового спама [14]. Существуют и не-текстовые приложения тематического моделирования в анализе изображений и видеопотоков [62, 82, 55, 145], в рекомендательных системах [173, 160, 77, 176, 175], в популяционной генетике [121], в биоинформатике для анализа нуклеотидных [74] и аминокислотных последовательностей [134, 72]. Другие приложения тематических моделей упоминаются в обзорах [45, 29, 36, 125].

Построение тематической модели по коллекции документов является некорректно поставленной оптимизационной задачей, которая может иметь бесконечное множество решений. Согласно теории регуляризации А. Н. Тихонова [15], решение такой задачи возможно доопределить и сделать устойчивым. Для этого к оптимизационному критерию добавляется *регуляризатор* — дополнительный критерий, учитывающий специфические особенности прикладной задачи или знания предметной области. В сложных приложениях дополнительных критериев может быть несколько.

Аддитивная регуляризация тематических моделей (additive regularization of topic models, ARTM) — это многокритериальный подход, в котором модель оптимизируется по взвешенной сумме критериев [4, 151]. ARTM позволяет строить модели с требуемыми свойствами, перенося регуляризаторы из одних моделей в другие или

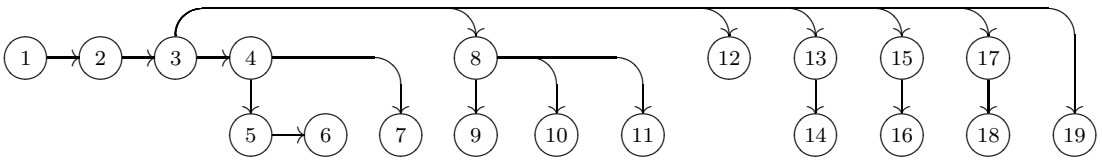
объединяя регуляризаторы от различных моделей. Для обучения любых моделей и их комбинаций используется один и тот же алгоритм, к которому регуляризаторы подключаются как модули [149, 56, 70]. Модульная технология реализована в библиотеке тематического моделирования с открытым кодом `BigARTM`, `http://bigartm.org` [149, 56]. Подчеркнём, что ARTM не является ещё одной моделью или методом — это общий подход к построению и комбинированию тематических моделей.

Доминирующим подходом к тематическому моделированию в настоящее время является байесовское обучение. В отличие от ARTM, в нём нет естественного разделения моделей на универсальный алгоритм и отторгаемые от него модули-регуляризаторы. Для каждой модели приходится заново проводить математический вывод и программную реализацию. Из-за сложности математического аппарата в статьях часто опускаются важные для понимания детали. Иногда авторы ограничиваются упрощённым описанием модели в виде порождающего процесса (generative story) или графической нотации (plate notation), однако последующий переход к алгоритму и его реализации остаётся неоднозначным и неочевидным. Эти барьеры препятствуют широкому распространению тематического моделирования: в индустрии анализа текстов редко можно встретить примеры использования тематических моделей сложнее морально устаревшей LDA (Latent Dirichlet Allocation) [33].

Основная цель данного обзора — показать разнообразие тематических моделей, сосредоточившись на важнейшем этапе моделирования — формализации постановки задачи. Тематическое моделирование обладает огромным запасом гибкости, позволяющим обрабатывать сложно структурированные данные и применять тематический анализ совместно с другими методами анализа текстов. Вторая цель — показать, что 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 доступнее и короче, без «заметания под ковёр» сложной математики. Сопоставимый по охвату и обстоятельности обзор байесовских моделей занял бы несколько сотен страниц.

Разделы 2–4 являются базовыми. В разделах 5–6 излагается теория байесовского подхода; она не обязательна для понимания последующего материала, но может быть полезна при чтении научной литературы. В разделах 7–14 в терминах регуляризации описываются различные виды тематических моделей. Они практически не связаны друг с другом, их можно читать в произвольном порядке или использовать как путеводитель по литературе. Разделы 15–16 описывают критерии качества тематических моделей. В разделе 17 рассматриваются особенности реализации алгоритмов, в том числе приёмы ускорения сходимости и распараллеливания. Раздел 18 содержит начальные сведения о библиотеке `BigARTM`. В разделе 19 обсуждается применение тематического моделирования для разведочного информационного поиска. В разделе 20 находится краткое заключение.

Ниже приведена схема зависимости разделов. Разделы верхнего ряда являются основными, нижнего — дополнительными.



2 Основы тематического моделирования

В этом разделе мы введём основные понятия и поставим задачу тематического моделирования как задачу приближённого низкорангового стохастического матричного разложения. С помощью формулы Байеса и частотных оценок условных вероятностей получим итерационный процесс, называемый EM-алгоритмом. Почему он действительно решает поставленную задачу, узнаем в следующем разделе.

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 [12] или методов автоматического выделения терминов (automatic term extraction, ATE), не требующих привлечения экспертов [52, 84, 131].

Распознавание именованных сущностей (named entities recognition, NER). Это названия объектов реального мира, относящихся к определённым категориям: персоны, организации, геолокации, события, даты, и т. д. Для распознавания именованных сущностей используются различные методы машинного обучения [103, 75, 109].

Пусть D — множество (коллекция) текстовых документов, W — 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 $\Omega_n = \big\{(w_i, d_i, t_i) \;\big|\; i = 1, \ldots, n\big\}$. 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 остаётся 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 $p(w\,|\,t)$:

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

Вероятностная тематическая модель порождения текста. Согласно формуле полной вероятности и гипотезе условной независимости, распределение термов в документе $p(w\,|\,d)$ описывается *вероятностной смесью* распределений термов в темах $\varphi_{wt} = p(w\,|\,t)$ с весами $\theta_{td} = p(t\,|\,d)$:

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

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.

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 $p(w|t_i)$;

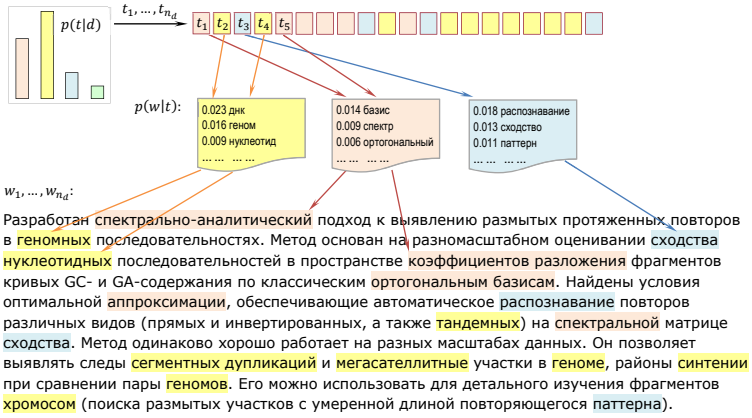


Рис. 1: Процесс порождения текстовой коллекции вероятностной тематической моделью (2): в каждой позиции i документа d_i сначала порождается тема $t_i \sim p(t|d_i)$, затем терм $w_i \sim p(w|t_i)$.

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 $\hat{p}(w|d) = \frac{n_{dw}}{n_d}$.

Распределение вида $p(t|x)$ будем называть *тематикой* объекта x. Можно говорить о тематике документа $p(t|d)$, терма $p(t|w)$, терма в документе $p(t|d,w)$.

Целью тематического моделирования является выявление тематической кластерной структуры текстовой коллекции, определение тематики документов и связанных с ними объектов, описание семантики каждой темы t словами естественного языка с помощью распределений $p(w|t)$.

Низкоранговое матричное разложение. Равенство (2) можно переписать в матричном виде. В левой части равенства находится матрица частот термов в документах $F = \big(\hat{p}(w|d)\big)_{W \times D}$, которая нам известна. Правая часть представляет собой произведение двух неизвестных матриц — *матрицы термов тем* $\Phi = (\varphi_{wt})_{W \times T}$ и *матрицы тем документов* $\Theta = (\theta_{td})_{T \times D}$. Обычно число тем $|T|$ много меньше $|D|$ и $|W|$, поэтому задача тематического моделирования сводится к поиску приближённого матричного разложения $F \approx \Phi\Theta$, ранг которого не превышает $|T|$.

Все три матрицы F, Φ, Θ являются *стохастическими*, то есть имеют неотрицательные нормированные столбцы f_d, φ_t, θ_d, представляющие дискретные распределения. Произведение $\Phi\Theta$ называется *стохастическим матричным разложением*.

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 частотные оценки не могут быть вычислены непосредственно по исходным данным, так как темы t_i неизвестны.

Согласно закону больших чисел, при $n \to \infty$ частотные оценки, определяемые формулами (3)–(4), 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:

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

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 её решения удобно применять метод простых итераций: сначала выбираются начальные приближения параметров φ_{wt} и θ_{td}, по ним вычисляются вспомогательные переменные p_{tdw}, которые позволяют найти следующее приближение параметров φ_{wt} и θ_{td}. Вычисления по формулам (6)–(8) продолжаются в цикле до сходимости.

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

Вход: коллекция D, число тем $|T|$, начальные приближения матриц Φ и Θ;
Выход: параметры модели Φ и Θ;

повторять
  обнулить n_{wt}, n_{td}, n_t для всех $d \in D,\ w \in W,\ t \in T$;
  для всех $d \in D,\ w \in d$
    $n_{tdw} := n_{dw}\varphi_{wt}\theta_{td} \,/\, \sum_\tau \varphi_{w\tau}\theta_{\tau d}$ для всех $t \in T$;
    увеличить n_{wt}, n_{td}, n_t на n_{tdw} для всех $t \in T$;
  $\varphi_{wt} := n_{wt}/n_t$ для всех $w \in W,\ t \in T$;
  $\theta_{td} := n_{td}/n_d$ для всех $d \in D,\ t \in T$;
пока Φ и Θ не сойдутся;

Этот итерационный процесс является частным случаем ЕМ-алгоритма, предназначенного для построения вероятностных моделей со скрытыми переменными [47]. Вычисление условных распределений скрытых переменных (6) называется Е-шагом (expectation), вычисление параметров модели (7)–(8) — М-шагом (maximization).

Далее мы выведем ЕМ-алгоритм из общей оптимизационной постановки задачи. Сейчас мы пришли к нему элементарным путём, который даёт простое интуитивное понимание сути ЕМ-алгоритма, 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 ЕМ-алгоритм. Вычисление переменных n_{wt}, n_{td}, n_t на М-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 Е-шаг можно встроить внутрь М-шага без дополнительных вычислительных затрат и без хранения трёхмерной матрицы p_{tdw}. Этот вариант реализации ЕМ-алгоритма будем называть *рациональным*; он показан в Алгоритме 2.

3 Аддитивная регуляризация

В этом разделе 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 разделах. Здесь мы поставим оптимизационную задачу с ограничениями и докажем лемму о максимизации гладкой функции на единичных симплексах, применимость которой выходит далеко за рамки тематического моделирования. С помощью этой леммы все последующие ЕМ-подобные алгоритмы будут выводиться в пару строчек (или чуть больше — по числу матриц в модели).

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 log-правдоподобия при ограничениях неотрицательности и нормировки:

$$\sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \varphi_{wt} \theta_{td} \;\to \max_{\Phi, \Theta}; \tag{9}$$

$$\sum_{w \in W} \varphi_{wt} = 1; \quad \varphi_{wt} \geqslant 0; \qquad \sum_{t \in T} \theta_{td} = 1; \quad \theta_{td} \geqslant 0. \tag{10}$$

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* [15]. 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) [4] основана на максимизации линейной комбинации логарифма правдоподобия и *регуляризаторов* $R_i(\Phi, \Theta)$ с неотрицательными *коэффициентами регуляризации* τ_i, $i = 1, \ldots, k$:

$$\sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \varphi_{wt} \theta_{td} + R(\Phi, \Theta) \;\to \max_{\Phi, \Theta}; \qquad R(\Phi, \Theta) = \sum_{i=1}^{k} \tau_i R_i(\Phi, \Theta); \tag{11}$$

при ограничениях неотрицательности и нормировки (10).

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

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:*

$$\begin{cases} \dfrac{\partial \mathscr{L}}{\partial x} = 0, \quad \mathscr{L}(x; \mu, \lambda) = f(x) - \sum\limits_{i=1}^{m} \mu_i g_i(x) - \sum\limits_{j=1}^{k} \lambda_j h_j(x); \\ g_i(x) \leqslant 0;\ h_j(x) = 0;\ \textit{(исходные ограничения)} \\ \mu_i \geqslant 0;\ \textit{(двойственные ограничения)} \\ \mu_i g_i(x) = 0;\ \textit{(условие дополняющей нежёсткости)} \end{cases} \tag{12}$$

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

Задача (11), (10) относится к классу невыпуклых задач математического программирования. Для неё возможно найти лишь локальный экстремум, качество которого будет зависеть от начального приближения. Однако заметим, что в ARTM варьирование самой постановки задачи путём подбора регуляризаторов $\tau_i R_i$ может влиять на решение сильнее, чем способ инициализации.

Лемма о максимизации 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:

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

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 симплексах). *Пусть функция $Q(\Omega)$ 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*

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

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

$$\omega_{ij} = \underset{i \in I_j}{\operatorname{norm}} \left(\omega_{ij} \frac{\partial Q}{\partial \omega_{ij}} \right), \quad i \in I_j; \tag{13}$$

при условии $(\forall i \in I_j)\ \omega_{ij} \frac{\partial Q}{\partial \omega_{ij}} \leqslant 0$ и $(\exists i \in I_j)\ \omega_{ij} \frac{\partial Q}{\partial \omega_{ij}} < 0$ — уравнениям

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

в противном случае — однородным уравнениям $\omega_{ij} \frac{\partial Q}{\partial \omega_{ij}} = 0$, $i \in I_j$.

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

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

где множители λ_j соответствуют ограничениям нормировки, μ_{ij} — ограничениям неотрицательности. Запишем условия Каруша–Куна–Таккера (12), приравняв нулю производные лагранжиана по параметрам модели:

$$\frac{\partial\mathscr{L}}{\partial\omega_{ij}} = \frac{\partial Q}{\partial\omega_{ij}} - \lambda_i + \mu_{ij} = 0; \quad \mu_{ij}\omega_{ij} = 0. \tag{14}$$

Зафиксируем j из J. Предположим, что $\omega_{ij} > 0$. Тогда $\mu_{ij} = 0$. Умножив обе части равенства (14) на ω_{ij}, получим уравнение

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

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 (13).

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 $\omega_{ij}\frac{\partial Q}{\partial\omega_{ij}} = 0$ при условиях неотрицательности и нормировки.

Лемма доказана.

Значение леммы о максимизации на единичных симплексах выходит далеко за пределы тематического моделирования. Она позволяет строить любые модели, параметрами которых являются дискретные вероятностные распределения.

В тематическом моделировании мы никогда не будем рассматривать второй и третий случай леммы 3.2, считая их вырожденными. Если в формуле *основного решения* (13) нормировочный знаменатель окажется равным нулю, то соответствующий вектор ω_j будем полагать нулевым и исключать его из модели, сокращая размерность пространства параметров и считая это полезным эффектом регуляризации.

Для решения системы уравнений удобно использовать метод простой итерации. Сначала выбирается начальное приближение для всех векторов ω_j, затем их значения многократно обновляются по формуле (13). Очерёдность и периодичность обновлений могут влиять на сходимость процесса и конечный результат. Заметим, что обновления похожи на градиентный метод максимизации $\omega_{ij} = \omega_{ij} + \eta\frac{\partial Q}{\partial\omega_{ij}}$, который не учитывает ограничений неотрицательности и нормировки. Кроме того, в нашем случае не возникает проблем с выбором градиентного шага η.

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 3.3. *Пусть функция $R(\Phi, \Theta)$ непрерывно дифференцируема. Тогда точка (Φ, Θ) локального экстремума задачи (11) с ограничениями (10) удовлетворяет системе уравнений со вспомогательными переменными $p_{tdw} = p(t\,|\,d, w)$, если из решения исключить нулевые столбцы матриц Φ, Θ:*

$$p_{tdw} = \underset{t\in T}{\operatorname{norm}}\big(\varphi_{wt}\theta_{td}\big); \tag{15}$$

$$\varphi_{wt} = \underset{w\in W}{\operatorname{norm}}\left(n_{wt} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}}\right); \qquad n_{wt} = \sum_{d\in D} n_{dw}p_{tdw}; \tag{16}$$

$$\theta_{td} = \underset{t\in T}{\operatorname{norm}}\left(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right); \qquad n_{td} = \sum_{w\in d} n_{dw}p_{tdw}. \tag{17}$$

Доказательство. Воспользуемся леммой 3.2 о максимизации на единичных симплексах, выделив вспомогательные переменные p_{tdw}, определённые в (15):

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

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

Теорема доказана.

Условия вырожденности. Рассмотрим условия, при которых формулы M-шага (16)–(17) могут давать вырожденные нулевые столбцы в матрицах Φ и Θ.

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 регуляризатору выгодно исключить данную тему из модели. Сокращение числа тем может быть желательным побочным эффектом регуляризации.

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 регуляризатору выгодно исключить данный документ из коллекции.

На практике вырожденность возникает редко. Если она нежелательна, то её нетрудно избежать. При постепенном уменьшении коэффициента регуляризации наступает момент, когда условие вырожденности перестаёт выполняться хотя бы для одного терма в теме (или хотя бы для одной темы в документе), и нулевой столбец в матрице решения переходит в ненулевой, удовлетворяющий условиям основного решения (13) в лемме о максимизации на симплексах.

Модель PLSA или вероятностного латентного семантического анализа (probabilistic latent semantic analysis) исторически является первой вероятностной тематической моделью. Она была предложена Томасом Хофманном в 1999 году [61]. В ARTM она соответствует нулевому регуляризатору, $R(\Phi, \Theta) = 0$. В этом случае система (15)–(17) совпадает с системой (6)–(8), которую мы получили ранее из элементарных соображений. Добавление регуляризации не меняет формулу E-шага.

В модели PLSA не может быть вырожденных тем или документов, поскольку вырожденность связана с отрицательностью производных регуляризатора.

Регуляризованный EM-алгоритм является применением метода простых итераций для решения системы (15)–(17). Сначала выбираются начальные приближения $\varphi_{wt}, \theta_{td}$, затем в цикле до сходимости чередуются *E-шаг* (15) и *M-шаг* (16)–(17). Рациональный вариант регуляризованного EM-алгоритма строится аналогично алгоритму 2, только шаги 6 и 7 заменяются формулами M-шага (16)–(17).

Известно, что EM-алгоритм без регуляризации сходится в слабом смысле: на каждой итерации правдоподобие увеличивается [47]. Аналогичные условия слабой сходимости для ARTM получены в [11].

Реализации EM-алгоритма могут различаться частотой обновления параметров модели φ_{wt} и θ_{td} по переменным n_{wt} и n_{td}. Частые обновления повышают скорость сходимости, но почти не влияют на значение правдоподобия в конце итераций [6].

Онлайновый EM-алгоритм считается наиболее быстрым и хорошо распараллеливается [60, 26]. Основная его идея состоит в том, что на большой коллекции матрица Φ может сойтись и перестать меняться задолго до окончания первой итерации. В таких случаях одного прохода по коллекции достаточно для построения модели. Поэтому онлайновые алгоритмы способны обрабатывать потоковые данные.

Детали параллельной реализации оффлайнового и онлайнового EM-алгоритма в библиотеке `BigARTM` описаны в разделах 17 и 18, ещё подробнее — в статьях [56, 3].

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. Одни регуляризаторы могут делать подготовительную работу для других. Некоторые регуляризаторы рекомендуется включать лишь после того, как ЕМ-алгоритм начал сходиться. Другие лучше отключать после того, как они выполнили свою работу. Некоторые регуляризаторы могут нейтрализовать друг друга, и тогда их приходится применять поочерёдно. 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 называются правила изменения коэффициентов регуляризации τ_i в ходе итераций ЕМ-алгоритма. Эти правила могут использовать текущие значения метрик качества и параметров модели.

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 М-шага (16) 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 (17) 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. При $\gamma_i = 0$ коэффициенты не различаются по документам.

4 Вероятностная регуляризация и модель LDA

Модель латентного размещения Дирихле (latent Dirichlet allocation) LDA [33] 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 не только ввели эту модель, но и поставили тематическое моделирование на рельсы байесовского обучения.

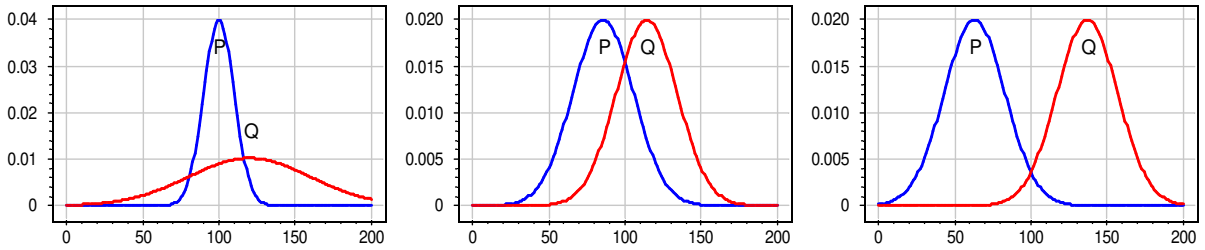
В данном разделе мы дадим более простое обоснование LDA с позиций аддитивной регуляризации. Для этого нам не понадобятся ни априорные распределения Дирихле, ни сложный математический аппарат байесовского вывода, рассмотрение которого мы отложим до раздела 6. Именно с этого простого определения мы и начнём, а оставшуюся часть раздела посвятим классическим обоснованиям.

Регуляризатор сглаживания и разреживания. Потребуем, чтобы столбцы φ_t матрицы Φ были похожи на заданный столбец $\beta = (\beta_w)$, а столбцы θ_d матрицы Θ были похожи на заданный столбец $\alpha = (\alpha_t)$. Формализуем наши требования с помощью суммы регуляризаторов максимума правдоподобия:

$$R(\Phi, \Theta) = \beta_0 \sum_{t\in T} \sum_{w\in W} \beta_w \ln \varphi_{wt} + \alpha_0 \sum_{d\in D} \sum_{t\in T} \alpha_t \ln \theta_{td} \to \max,$$

где β_0 и α_0 — коэффициенты регуляризации. Подставим критерий регуляризации $R(\Phi, \Theta)$ в формулы М-шага (16)–(17):

$$\varphi_{wt} = \underset{w\in W}{\operatorname{norm}}\big(n_{wt} + \beta_0\beta_w\big); \qquad \theta_{td} = \underset{t\in T}{\operatorname{norm}}\big(n_{td} + \alpha_0\alpha_t\big).$$



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

Рис. 2: 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 приблизительно одинакова на левом и среднем графиках, вложенность Q в P — на левом и правом графиках.

Это простейшая модификация М-шага и единственная из возможных, в которой поправки частотных оценок условных вероятностей φ_{wt} и θ_{td} являются константами, а не функциями от φ_{wt} и θ_{td}. Чем больше значения коэффициентов β_0 и α_0, тем сильнее столбцы φ_t и θ_d будут похожи на векторы β и α соответственно. Этот эффект будем называть *сглаживанием* частотных оценок условных вероятностей.

Значения коэффициентов β_0 и α_0 могут быть и отрицательными. В таком случае регуляризатор соответствует требованию, чтобы столбцы φ_t и θ_d были *не* похожи на β и α соответственно. Формула М-шага показывает, что это может приводить к обнулению тех условных вероятностей, которые и так были близки к нулю. Этот эффект будем называть *разреживанием* частотных оценок условных вероятностей.

Возможна эквивалентная запись регуляризатора через дивергенцию Кульбака–Лейблера, которая является мерой различности пары вероятностных распределений:

$$R(\Phi, \Theta) = -\beta_0 \sum_{t\in T} \mathrm{KL}(\beta_w \| \varphi_{wt}) - \alpha_0 \sum_{d\in D} \mathrm{KL}(\alpha_t \| \theta_{td}) \to \max.$$

Дивергенция Кульбака–Лейблера (*KL-дивергенция*, относительная энтропия) далее будет одним из важнейших инструментов конструирования регуляризаторов. Это несимметричная функция расстояния между дискретными распределениями $P = (p_i)_{i=1}^n$ и $Q = (q_i)_{i=1}^n$, с совпадающими носителями, $\{i\colon p_i > 0\} = \{i\colon q_i > 0\}$:

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

где $H(P) = -\sum_i p_i \ln p_i$ — *энтропия* распределения P, $H(P,Q) = -\sum_i p_i \ln q_i$ — *кросс-энтропия* распределений P и Q. Обозначение KL_i не является общепринятым, но оно удобно, когда надо показать, по какому индексу производится суммирование.

KL-дивергенция неотрицательна и равна нулю тогда и только тогда, когда $P = Q$.

Если $\mathrm{KL}(P\|Q) < \mathrm{KL}(Q\|P)$, то распределение P сильнее вложено в Q, чем Q в P, см. рис. 2. Таким образом, KL является мерой вложенности двух распределений.

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

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

В частности, заметим, что максимизация правдоподобия для тематического моделирования (9) эквивалентна минимизации взвешенной суммы KL-дивергенций 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 функционала качества может быть полезна при моделировании коллекций, содержащих документы одинаковой важности, но существенно разной длины.

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 (9), в которой логарифм априорного распределения становится регуляризатором:

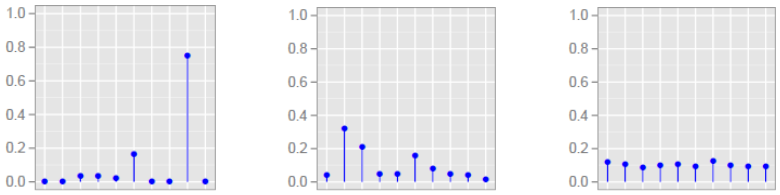
$$\sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{t\in T} \varphi_{wt}\theta_{td} + \underbrace{\ln p(\Phi, \Theta; \gamma)}_{R(\Phi,\Theta)} \;\to\; \max_{\Phi,\Theta,\gamma}. \tag{18}$$

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. Однако если решать задачу оптимизации гиперпараметров γ, то его уже придётся учитывать.

Априорные распределения Дирихле. Основной мотивацией для введения модели LDA в [33] 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 [33] было введено предположение, что столбцы этих матриц являются случайными векторами и порождаются распределениями Дирихле с гиперпараметрами $\alpha \in \mathbb{R}^T$ и $\beta \in \mathbb{R}^W$ соответственно:

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

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



$\mathrm{Dir}(\varphi; 0.1)$ $\mathrm{Dir}(\varphi; 1)$ (равномерное) $\mathrm{Dir}(\varphi; 100)$

Рис. 3: Пример трёх неотрицательных нормированных векторов φ_t размерности $|W| = 10$, порождённых соответственно тремя симметричными распределениями Дирихле с параметрами 0.1, 1, 100.

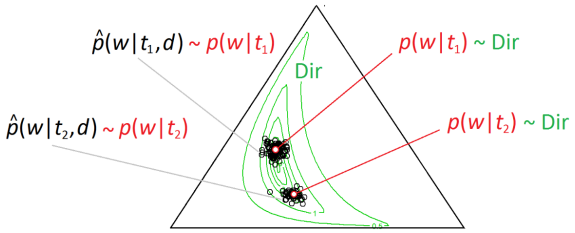


Рис. 4: Распределение $\mathrm{Dir}(\varphi|\alpha)$ порождает векторы тем $\varphi_t = p(w|t)$, которые порождают мультиномиальные распределения $\hat{p}(w|t,d)$ на единичном симплексе в пространстве размерности $|W| = 3$.

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

Использование распределений Дирихле мотивировано тремя причинами.

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.

Аналогично, гипотеза о разреженности распределений $\theta_{td} = p(t|d)$ формализует 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 $\varphi_{wt} = p(w|t)$, 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 разделах. Именно математическое удобство предопре-

делило популярность распределения Дирихле и модели LDA в тематическом моделировании, хотя убедительных лингвистических обоснований оно не имеет.

Согласно (18), модели LDA соответствует регуляризатор, с точностью до константы равный сумме логарифмов априорных распределений Дирихле:

$$R(\Phi,\Theta) = \ln \prod_{t\in T} \mathrm{Dir}(\varphi_t;\beta) \prod_{d\in D} \mathrm{Dir}(\theta_d;\alpha) + \mathrm{const} = \\ = \sum_{t\in T}\sum_{w\in W} (\beta_w - 1)\ln \varphi_{wt} + \sum_{d\in D}\sum_{t\in T} (\alpha_t - 1)\ln \theta_{td}. \tag{19}$$

Применение уравнений (16)–(17) к этому регуляризатору даёт формулы M-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 M-шага переходят в несмещённые частотные оценки условных вероятностей, а модель LDA переходит в PLSA [57].

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 $0 < \beta_w < 1,\ 0 < \alpha_t < 1$ регуляризатор имеет разреживающий эффект и способен обнулять малые вероятности. Однако 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 модели LDA. Регуляризатор (19) можно эквивалентным образом записать через KL-дивергенции:

$$R(\Phi,\Theta) = |W| \sum_{t\in T} \mathrm{KL}_w\big(\tfrac{1}{|W|} \,\big\|\, \varphi_{wt}\big) - \beta_0 \sum_{t\in T} \mathrm{KL}_w\big(\tfrac{\beta_w}{\beta_0} \,\big\|\, \varphi_{wt}\big) + \\ + |T| \sum_{d\in D} \mathrm{KL}_t\big(\tfrac{1}{|T|} \,\big\|\, \theta_{td}\big) - \alpha_0 \sum_{d\in D} \mathrm{KL}_t\big(\tfrac{\alpha_t}{\alpha_0} \,\big\|\, \theta_{td}\big).$$

Отсюда следует, что модель LDA оказывает сглаживающие и разреживающие воздействия на матрицы Φ, Θ. Все столбцы матрицы Φ должны быть близки к одному и тому же распределению $\frac{\beta_w}{\beta_0}$, причём параметр β_0 становится коэффициентом регуляризации. Аналогично, все столбцы матрицы Θ должны быть близки к распределению $\frac{\alpha_t}{\alpha_0}$, и этим требованием управляет коэффициент регуляризации α_0. Кроме этих сглаживающих воздействий имеются слабые неуправляемые разреживающие воздействия: столбцы обеих матриц должны быть далеки от равномерного распределения. Наиболее удалены от равномерного вырожденные распределения, в которых единичная вероятность сконцентрирована в единственном элементе. Поэтому разреживание приводит к обнулению малых вероятностей в матрицах Φ, Θ.

Таким образом, регуляризатор сглаживания и разреживания обобщает модель LDA, отменяя избыточное ограничение неотрицательности гиперпараметров. Кроме того, принципы максимума правдоподобия или минимума KL-дивергенции представляются более простыми и удобными инструментами введения дополнительных ограничений на модель, чем априорные распределения Дирихле.

5 Теория EM-алгоритма

В этом разделе мы погрузимся в теорию EM-алгоритма. Рассмотрим более общий классический способ его вывода, который позволяет обосновать сходимость. Заодно обогатим его возможностью аддитивной регуляризации. Материал этого раздела понадобится только в следующем разделе, ещё более теоретическом. Если нетерпится поскорее узнать о практических аспектах тематического моделирования, то оба раздела можно целиком пропустить.

Общий EM-алгоритм с регуляризацией. Классический EM-алгоритм [47] предназначен для построения широкого класса вероятностных порождающих моделей со скрытыми переменными. Тематическое моделирование для него — лишь частный случай. 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 $R(\Omega)$:

$$\ln p(X \,|\, \Omega) + R(\Omega) = \ln \sum_Z p(X, Z \,|\, \Omega) + R(\Omega) \;\to\; \max_\Omega. \tag{20}$$

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 5.1. *Если функционал* (20) *достаточно гладкий, то точка* Ω *его локального максимума удовлетворяет системе уравнений, решение которой методом простых итераций сводится к чередованию двух шагов:*

$$\textit{E-шаг:}\quad q(Z) = p(Z \,|\, X, \Omega); \tag{21}$$

$$\textit{M-шаг:}\quad \sum_Z q(Z) \ln p(X, Z \,|\, \Omega) + R(\Omega) \;\to\; \max_\Omega. \tag{22}$$

Доказательство. Запишем необходимые условия локального экстремума для задачи максимизации гладкого функционала (20):

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

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 суммирования:

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

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

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

Полученное уравнение является необходимым условием локального экстремума задачи М-шага (22), если фиксировать распределение $p(Z \mid X, \Omega)$ так, чтобы оно не зависело от параметра Ω. Именно это и достигается вычислением $q(Z)$ на Е-шаге (21) и последующей его подстановкой в (22). Таким образом, мы получили систему уравнений, эквивалентную необходимым условиям максимума функционала (20).

Теорема доказана.

Следствие 5.2. *Если задача оптимизации* (20) *имеет ограничения* $g_i(\Omega) \leqslant 0$, $h_j(\Omega) = 0$, *то система уравнений* (21)–(22) *остаётся в силе, при этом задача оптимизации* (22) *имеет те же ограничения.*

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

Теорема 5.3. *В итерационном процессе* (21)–(22) *значение функционала не уменьшается на каждом шаге.*

Доказательство. Введём произвольное распределение $q(Z)$. Воспользуемся условием нормировки $\sum_Z q(Z) = 1$ и тождеством $p(X \mid \Omega) = \frac{p(X,Z \mid \Omega)}{p(Z \mid X,\Omega)}$:

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

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

$$\ln p(X \mid \Omega) = \underbrace{\sum_Z q(Z) \ln \frac{p(X, Z \mid \Omega)}{q(Z)}}_{L(q,\Omega)} + \underbrace{\sum_Z q(Z) \ln \frac{q(Z)}{p(Z \mid X, \Omega)}}_{\mathrm{KL}(q(Z) \| p(Z \mid X, \Omega)) \geqslant 0}. \tag{23}$$

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 $\ln p(X \mid \Omega)$.

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

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

Максимизация $L(q, \Omega)$ по q эквивалентна минимизации $\mathrm{KL}\big(q(Z) \,\big\|\, p(Z \mid X, \Omega)\big)$ по q, поскольку их сумма (23) не зависит от q. Минимум KL-дивергенции равен нулю и достигается при $q(Z) = p(Z \mid X, \Omega)$, что совпадает с формулой Е-шага (21). Обнуление KL-дивергенции в (23) означает, что $L(q, \Omega)$ является достигаемой нижней оценкой для $\ln p(X \mid \Omega)$, а задача максимизации на М-шаге совпадает с (22).

Таким образом, EM-алгоритм (21)–(22) 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.

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

Общий ЕМ-алгоритм для ARTM. Применим ЕМ-алгоритм (21)–(22) к задаче тематического моделирования, вернувшись к исходным обозначениям: $X = (d_i, w_i)_{i=1}^n$, $Z = (t_i)_{i=1}^n$, $\Omega = (\Phi, \Theta)$.

Следующая лемма показывает, что в случае тематического моделирования сумма по всевозможным Z в задаче М-шага (22) существенно упрощается.

Лемма 5.4. *Точка (Φ, Θ) локального максимума регуляризованного log-правдоподобия в задаче тематического моделирования (11) удовлетворяет системе уравнений, решение которой методом простых итераций сводится к чередованию двух шагов:*

$$\text{Е-шаг:} \quad p(t\,|\,d,w) = \operatorname*{norm}_{t\in T}\big(\varphi_{wt}\theta_{td}\big), \quad \text{для всех } d\in D,\ w\in W,\ t\in T; \tag{24}$$

$$\text{М-шаг:} \quad \sum_{d\in D}\sum_{w\in W}\sum_{t\in T} n_{dw}p(t\,|\,d,w)\ln\big(\varphi_{wt}\theta_{td}\big) + R(\Phi,\Theta) \;\to\; \max_{\Phi,\Theta}. \tag{25}$$

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

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

Подставим в общую формулу М-шага (22) 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:

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

Лемма доказана.

Задача (25) является максимизацией регуляризованного правдоподобия для вероятностной модели $p(w,t\,|\,d)$, которая настраивается по эмпирическим частотам $n_{tdw} = n_{dw}p(t\,|\,d,w)$. Это полезная промежуточная форма постановки задачи тематического моделирования при известных распределениях скрытых переменных.

Покажем, что от неё легко перейти к EM-алгоритму для ARTM (15)–(17). Воспользуемся прежними обозначениями n_{wt} из (7) и n_{td} из (8). Тогда критерий (25) раскладывается в сумму двух слагаемых и регуляризатора; первое зависит только от Φ, второе — только от Θ:

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

Применяя к данному критерию лемму 3.2 о максимизации на единичных симплексах, немедленно получаем формулы M-шага для ARTM (16)–(17). В совокупности такой способ вывода длиннее, но его преимущество в том, что попутно удаётся обосновать сходимость EM-алгоритма.

6 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 условных вероятностей в матрицах Φ и Θ, а не их распределения.

Остаётся лишь гадать, 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. Возможно, тематическое моделирование перескочило через 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 применяться методы байесовского вывода. Тематическое моделирование оказалась относительно новой задачей в тот момент, когда байесовское обучение было на пике популярности. Простота и богатые возможности обычной, не байесовской, регуляризации оказались незамеченными. Это упущение как раз и призван устранить подход ARTM.

Несмотря на высказанные выше соображения, байесовское обучение крайне важно для тематического моделирования хотя бы потому, что большинство публикаций написано на этом математическом языке.

Мы рассмотрим два наиболее популярных подхода: вариационный байесовский вывод (variational Bayes, VB) [143] и сэмплирование Гиббса (Gibbs sampling, GS) [139]. 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 данном разделе, далее использоваться не будет. Байесовские тематические модели, как правило, удаётся переформулировать в терминах регуляризации намного яснее и проще. Примеров таких «дебайесенизированных» моделей будет много в следующих разделах.

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 $\propto$ 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 предыдущем разделе. Тогда задача сводится к максимизации логарифма правдоподобия с вероятностным регуляризатором:

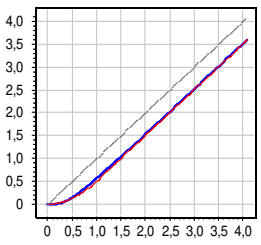
$$\ln p(X\,|\,\Omega) + \ln p(\Omega\,|\,\gamma) \;\to\; \max_{\Omega,\gamma}.$$

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. Такой подход даёт больше информации о параметрах модели, позволяет строить для них доверительные и интервальные оценки.

Свойства распределения Дирихле. Модель LDA основана на распределениях Дирихле, поэтому приведём в справочном порядке некоторые его свойства:

- $\mathsf{E}\theta_t = \int \theta_t \operatorname{Dir}(\theta|\alpha)\,d\theta = \frac{\alpha_t}{\alpha_0} = \operatorname*{norm}_t(\alpha_t)$ — математическое ожидание θ_t;
- $\hat{\theta}_t = \frac{\alpha_t - 1}{\alpha_0 - T} = \operatorname*{norm}_t(\alpha_t - 1)$ — 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 6.1. *Пусть задано разбиение множества переменных Y на блоки Y_j, $j \in J$. Тогда решение задачи* $\mathrm{KL}\big(q(Y) \,\big\|\, p(Y \,|\, X, \gamma)\big) \to \min\limits_{q}$ *в семействе факторизованных распределений* $q(Y) = \prod\limits_{j \in J} q_j(Y_j)$ *удовлетворяет системе уравнений*

$$\ln q_j(Y_j) = \mathsf{E}_{q \setminus j} \ln p(X, Y \,|\, \gamma) + \mathrm{const}, \quad j \in J, \tag{26}$$

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

Доказательство можно найти, например, в [27].

Для решения системы (26) можно использовать метод простой итерации.

Применим эту теорему к нашему случаю, когда $Y = (Z, \Phi, \Theta)$, $\gamma = (\alpha, \beta)$. Будем приближать 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 $J = \{1, \ldots, n\} \sqcup T \sqcup D$ — 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 (26), распишем логарифм распределения $p(X, Z, \Phi, \Theta \,|\, \alpha, \beta)$, переводя слагаемые, не зависящие от переменных t_i, φ_t, θ_d, в const:

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

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 (26) относительно $q_t(\varphi_t)$:

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

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 (26) относительно $q_d(\theta_d)$:

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

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 $\theta_{td} = \operatorname*{norm}_t(\tilde{\alpha}_{td})$.

Наконец, рассмотрим уравнение (26) 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:

$$q_i(t) = \operatorname*{norm}_{t\in T} \left(\frac{E\big(n_{w_i t} + \beta_{w_i}\big)}{E\big(\sum_w (n_{wt} + \beta_w)\big)} \cdot \frac{E\big(n_{td_i} + \alpha_t\big)}{E\big(\sum_t (n_{td_i} + \alpha_t)\big)} \right), \tag{27}$$

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 (15) 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 (16)–(17), 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 (26) 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 через математическое ожидание 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 [24].

Сэмплирование Гиббса. Идея этого подхода в том, чтобы сначала оценить скрытые переменные с помощью 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.

Разберёмся сначала с апостериорным распределением. Пусть $p(\Phi, \Theta\,|\,\alpha, \beta)$ — априорное распределение Дирихле. Распределение Дирихле является сопряжённым к мультиномиальному, поэтому апостериорное также будет принадлежать семейству распределений Дирихле:

$$\begin{aligned}
&p(\Phi, \Theta\,|\,X, Z, \alpha, \beta) \propto p(\Phi, \Theta, X, Z\,|\,\alpha, \beta) \propto p(X, Z\,|\,\Phi, \Theta) p(\Phi, \Theta\,|\,\alpha, \beta) \\
&\propto \prod_{d,w,t} (\varphi_{wt}\theta_{td})^{n_{dwt}} \prod_{t\in T} \operatorname{Dir}(\varphi_t\,|\,\beta) \prod_{d\in D} \operatorname{Dir}(\theta_d\,|\,\alpha) \\
&\propto \prod_{t\in T}\prod_{d,w} \varphi_{wt}^{n_{dwt}} \varphi_{wt}^{\beta_w - 1} \prod_{d\in D}\prod_{w,t} \theta_{td}^{n_{dwt}} \theta_{td}^{\alpha_t - 1} \\
&\propto \prod_{t\in T}\prod_{w} \varphi_{wt}^{n_{wt} + \beta_w - 1} \prod_{d\in D}\prod_{t} \theta_{td}^{n_{td} + \alpha_t - 1} \\
&\propto \prod_{t\in T} \operatorname{Dir}(\varphi_t\,|\,\tilde{\beta}_t) \prod_{d\in D} \operatorname{Dir}(\theta_d\,|\,\tilde{\alpha}_d), \quad \tilde{\beta}_{wt} = n_{wt} + \beta_w, \quad \tilde{\alpha}_{td} = n_{td} + \alpha_t,
\end{aligned} \tag{28}$$

где счётчики n_{dwt}, n_{wt} и n_{td} определяются согласно (4) через значения скрытых переменных Z.

Сэмплирование случайного вектора Z из многомерного распределения основано на *теореме о сходимости процесса сэмплирования Гиббса.*

Теорема 6.2. *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 [27].

Чтобы воспользоваться этой теоремой, 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 (28), но лишь с точностью до нормировочных множителей. Теперь они нам понадобятся, поэтому разберёмся с ними аккуратнее:

$$\begin{aligned}
p(X, Z \,|\, \alpha, \beta) &= \int_\Phi \int_\Theta p(X, Z \,|\, \Phi, \Theta) p(\Phi, \Theta \,|\, \alpha, \beta) d\Phi d\Theta = \\
&= \int_\Phi \int_\Theta \prod_{w,t} \varphi_{wt}^{n_{wt}} \prod_{t,d} \theta_{td}^{n_{td}} \prod_d p_d^{n_d} \prod_{t \in T} \mathrm{Dir}(\varphi_t \,|\, \beta) \prod_{d \in D} \mathrm{Dir}(\theta_d \,|\, \alpha) d\Phi d\Theta = \\
&= \prod_{t \in T} \frac{\Gamma(\sum_w \beta_w)}{\prod_w \Gamma(\beta_w)} \int_{\varphi_t} \underbrace{\prod_w \varphi_{wt}^{\tilde\beta_{wt} - 1} d\varphi_t}_{\propto \mathrm{Dir}(\varphi_t \,|\, \tilde\beta_t)} \prod_{d \in D} p_d^{n_d} \frac{\Gamma(\sum_t \alpha_t)}{\prod_t \Gamma(\alpha_t)} \int_{\theta_d} \underbrace{\prod_t \theta_{td}^{\tilde\alpha_{td} - 1} d\theta_d}_{\propto \mathrm{Dir}(\theta_d \,|\, \tilde\alpha_d)} = \\
&= \prod_{t \in T} \frac{\Gamma(\sum_w \beta_w)}{\prod_w \Gamma(\beta_w)} \frac{\prod_w \Gamma(\tilde\beta_{wt})}{\Gamma(\sum_w \tilde\beta_{wt})} \prod_{d \in D} p_d^{n_d} \frac{\Gamma(\sum_t \alpha_t)}{\prod_t \Gamma(\alpha_t)} \frac{\prod_t \Gamma(\tilde\alpha_{td})}{\Gamma(\sum_t \tilde\alpha_{td})}.
\end{aligned} \tag{29}$$

Распределение $p(X, Z_{\setminus i} \,|\, \alpha, \beta)$ отличается от полученного выше $p(X, Z \,|\, \alpha, \beta)$ лишь тем, что оно строится по выборке 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 6.2, 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):

$$\begin{aligned}
p(t_i \,|\, X, Z_{\setminus i}, \alpha, \beta) &= \frac{p(X, Z \,|\, \alpha, \beta)}{p(X, Z_{\setminus i} \,|\, \alpha, \beta)} = \\
&= \frac{\Gamma(n_{w_i t_i} + \beta_{w_i}) \, \Gamma(\sum_w (n_{wt_i} + \beta_w) - 1)}{\Gamma(n_{w_i t_i} + \beta_{w_i} - 1) \, \Gamma(\sum_w (n_{wt_i} + \beta_w))} \frac{\Gamma(n_{t_i d_i} + \alpha_{t_i}) \, \Gamma(\sum_t (n_{td_i} + \alpha_t) - 1)}{\Gamma(n_{t_i d_i} + \alpha_{t_i} - 1) \, \Gamma(\sum_t (n_{td_i} + \alpha_t))}.
\end{aligned}$$

В этом выражении воспользуемся свойством гамма-функции $\frac{\Gamma(x)}{\Gamma(x-1)} = x - 1$:

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

По сути это оценка условной вероятности $p(t \,|\, d_i, w_i) = \mathrm{norm}_t(\varphi_{w_i t} \theta_{t d_i})$ на E-шаге EM-алгоритма для модели LDA. Таким образом, мы снова получаем EM-подобный

Алгоритм 3. 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** $t := t_i$; $--n_{wt}$; $--n_{td}$; $--n_t$; $--n_d$;
    $p(t\,|\,d,w) = \underset{t\in T}{\operatorname{norm}}\big(\frac{n_{wt}+\beta_w}{n_t+\beta_0} \cdot \frac{n_{td}+\alpha_t}{n_d+\alpha_0}\big)$ 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 $d \in D,\ t \in T$;

алгоритм, хотя задача максимизации правдоподобия даже не ставилась. Как и в случае с вариационными байесовским выводом, алгоритм отличается несущественными поправками к частотным оценкам 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 сэмплированным темам 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 (28):

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

В алгоритме 3 показана идея реализации, предложенная в [139]. 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 6.2. 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.

Сходство EM-подобных алгоритмов PLSA, MAP, VB, GS и ещё нескольких их вариантов было замечено в [24]. В работе [6] сэмплирование единственной темы из распределения $t \sim p(t\,|\,d,w)$ рассматривалось как отдельная эвристика, которую можно использовать в любом EM-подобном алгоритме тематического моделирования, начиная с PLSA. 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 [139], используются симметричные априорные распределения Дирихле с параметрами $\alpha_t = 50/|T|$, $\beta_w = 0.01$. Более тонкие исследования [156] показали, что лучше оптимизировать вектор $\alpha = (\alpha_1, \ldots, \alpha_T)$ в несимметричном распределении $\mathrm{Dir}(\theta_d \,|\, \alpha)$ и подбирать числовой параметр $\beta_w \ll 1$ в симметричном распределении $\mathrm{Dir}(\varphi_t \,|\, \beta)$.

Для оптимизации вектора α в [156] предлагается максимизировать правдоподобие $P(X, Z \,|\, \alpha, \beta)$ при фиксированных Z и β. Отбрасывая в (29) множители, не зависящие от α, получаем оптимизационную задачу:

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

В диссертации [155] сравнивалось более десятка численных методов её решения. Приведём лишь самый простой из них — метод неподвижной точки [101]. В нём вектор $\alpha = (\alpha_t)$ 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 всей коллекции.

Эксперименты в [156] 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. 5 показаны примеры зависимостей между случайными переменными. Наблюдаемые переменные изображаются закрашенным кружком. Выборка переменных, генерируемых одним распределением, изображается прямоугольником, рис. 6. Такие изображения будем называть *графической нотацией* (plate notation).

Альтернативной текстуальной формой представления вероятностной модели является *генеративная история* (generative story), в которой описывается алгоритм порождения данных. На рис. 7 показаны оба представления, графическое и текстуальное, для тематических моделей PLSA и LDA.

Подборка графических нотаций на рис. 8 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. В результате модель может восстанавливаться по картинке неоднозначно. Генеративная история лишена этого недостатка и даёт более точное описание модели. Графическая нотация и генеративная история

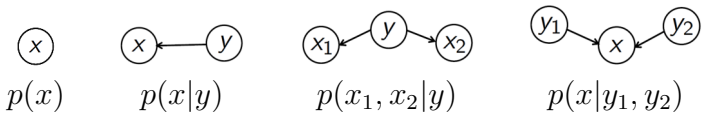


Рис. 5: Представление условных зависимостей в графической нотации.

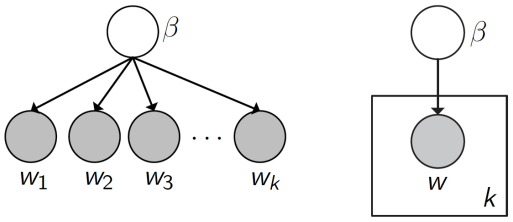
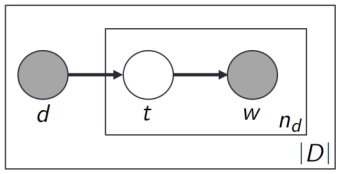


Рис. 6: Графическая нотация выборки $w_1, \ldots, w_k$, порождаемой распределением $\beta_w = p(w)$.

Модель PLSA:

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



Модель LDA:

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

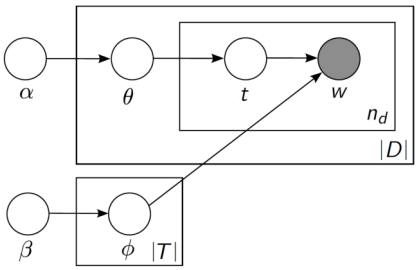


Рис. 7: Генеративная история и графическая нотация для моделей PLSA и LDA.

могут быть полезны для понимания общей идеи и конструкции модели. Однако их не достаточно, чтобы однозначно воспроизвести переход от модели к алгоритму её обучения. К сожалению, во многих публикациях авторы опускают этот переход для краткости изложения, что отнюдь не добавляет ясности таким работам.

Сравнение ARTM и байесовского подхода. К недостаткам байесовского вывода можно отнести техническую сложность добавления и комбинирования требований к модели в виде оптимизационных критериев. В нём нет удобных механизмов регуляризации, поскольку нет, собственно, и задачи оптимизации по (Φ, Θ). Чтобы учесть дополнительную информацию, приходится менять априорные распределения или всю структуру модели. Если априорные распределения не являются распределениями Дирихле, вывод становится заметно сложнее.

Распределение Дирихле оказывается «на особом положении» в байесовском подходе. Оно не имеет убедительных лингвистических обоснований, тем не менее, в литературе большинство моделей строятся с его использованием. Это объясняется его математическим удобством, а именно, свойством сопряжённости с мультиномиаль-

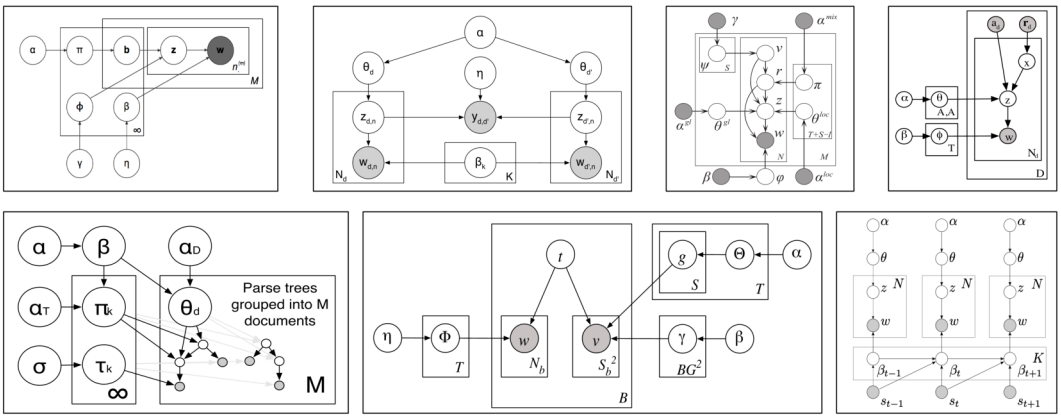


Рис. 8: Примеры графических нотаций из публикаций по тематическому моделированию.

ным распределением. В ARTM нет никаких оснований предпочитать распределение Дирихле другим регуляризаторам.

Постановка оптимизационной задачи вида (18) даёт возможность интерпретировать априорное распределение как вероятностный регуляризатор, отделить его от конкретной модели и использовать в других моделях. Аддитивность регуляризаторов приводит к модульной технологии тематического моделирования, которая реализована в проекте `BigARTM` [8, 149]. При решении прикладных задач комбинирование готовых регуляризаторов позволяет строить модели с заданными свойствами без дополнительных математических выкладок и программирования. Создание такой технологии в рамках байесовского подхода едва ли возможно.

В дальнейшем мы будем рассматривать все тематические модели в терминах регуляризации, даже если в исходных работах они строились байесовскими методами. Реформализация байесовских моделей через регуляризацию, как правило, не представляет особого труда, более того, приводит к радикальному упрощению изложения без существенных изменений в конечных алгоритмах. Мы не будем использовать графическую нотацию, поскольку в ARTM даже сложные модели формализуются настолько просто, что потребности в графических представлениях не возникает.

7 Модели сглаживания и разреживания

Отказ от априорных распределений Дирихле позволяет обобщить модель LDA: снять ограничения на знаки гиперпараметров в (19) 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 вероятностей φ_{wt} и θ_{td} равны нулю.

Разреженность представляется естественным необходимым условием интерпретируемости тематической модели. Кроме того, разреженные структуры данных и алгоритмов позволяют заметно сокращать время вычислений и расход памяти как в процессе построения модели, так и при дальнейшем её использовании.

Многие попытки разреживания модели LDA приводили к чрезмерно сложным конструкциям [133, 51, 159, 76, 40] из-за внутреннего противоречия между разреженностью и положительностью параметров в распределении Дирихле. Проблема

решается просто, если в регуляризаторе (19) позволить гиперпараметрам α_t, β_w принимать отрицательные значения. По всей видимости, впервые это было предложено в динамической модели PLSA для обработки видеопотоков [145], 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».

Обобщение LDA. Максимизация апостериорной вероятности в модели LDA приводит к регуляризатору (19). Снимем ограничения неотрицательности с параметров β_w, α_t и позволим задавать их индивидуально для каждой ячейки матриц Φ и Θ. Получим *обобщённый регуляризатор сглаживания и разреживания*:

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

Подставив этот регуляризатор в (16)–(17), получим формулы M-шага:

$$\varphi_{wt} = \underset{w \in W}{\operatorname{norm}}\big(n_{wt} + \beta_{wt}\big); \tag{30}$$

$$\theta_{td} = \underset{t \in T}{\operatorname{norm}}\big(n_{td} + \alpha_{td}\big). \tag{31}$$

Положительное значение параметра β_{wt} или α_{td} соответствует сглаживанию, отрицательное — разреживанию. При таком обобщении некорректно (да и нет необходимости) говорить об априорных распределениях Дирихле. Обобщённый регуляризатор оставляет свободу выбора ячеек матриц для разреживания и сглаживания.

Частичное обучение. В процессе создания, использования или оценивания тематической модели эксперты, пользователи или асессоры 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.

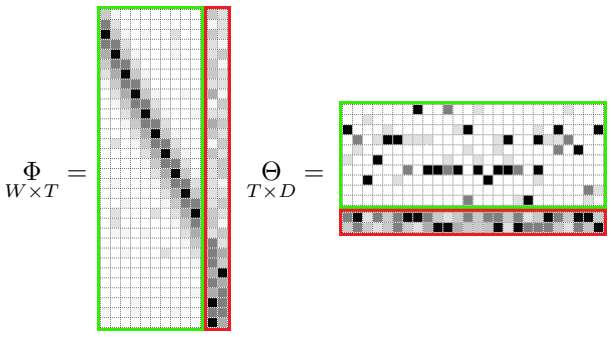


Рис. 9: Структура разреженности матриц Φ и Θ с предметными и фоновыми темами.

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. 9. 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 $p(d|t)$ также разрежены, так как каждая предметная тема присутствует в относительно небольшой доле документов.

Фоновые темы $t \in B$ содержат слова общей лексики, которых не должно быть в предметных темах. Их распределения $p(w|t)$ и $p(d|t)$ сглажены, так как эти слова присутствуют в большинстве документов. Тематическую модель с фоновыми темами можно рассматривать как обобщение робастных моделей [38, 120], 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 [20]. 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 «классифицировать иголки в стоге сена» [34]. 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 [66], поиска в социальных медиа информации, связанной с болезнями, симптомами и методами лечения [113, 114], с преступностью и экстремизмом [87, 132], с национальностями и межнациональными отношениями [34, 71, 111].

В модели ATAM (ailment topic aspects model) в качестве сглаживающего распределения β_{wt} использовалась большая коллекция медицинских статей [114].

В моделях SSLDA (semi-supervised LDA) и ISLDA (interval semi-supervised LDA) для поиска этно-релевантных тем в постах социальных сетей использовалось сгла-

живание по словарю из нескольких сотен этнонимов [34]. 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 конфликтов. Альтернативный подход заключается в том, чтобы задать число этно-релевантных тем и применить к ним общее сглаживание по словарю этнонимов. Тематическая модель сама определит, как разделить их по этничностям [21, 22]. Недостаток этого подхода в том, что интерпретируемость найденных тем приходится проверять вручную.

Декоррелирование. Тематическая модель не должна содержать дублирующихся или похожих тем. Чем различнее темы, тем информативнее модель. 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 М-шага, согласно (16), имеет вид

$$\varphi_{wt} = \underset{w \in W}{\operatorname{norm}} \Big(n_{wt} - \tau \varphi_{wt} \sum_{s \in T \setminus t} \varphi_{ws} \Big). \tag{32}$$

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 [141] 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 [7, 151, 150].

Декоррелирование впервые было предложено в модели TWC-LDA (topic-weak-correlated LDA) в рамках байесовского подхода [141]. Соответствующее априорное распределение не является сопряжённым к мультиномиальному, поэтому байесовский вывод сталкивается с техническими трудностями. В ARTM расчётная формула М-шага (32) 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 [7, 150, 151, 152, 17]. Подобрав коэффициенты регуляризации, можно одновременно значительно улучшить разреженность, контрастность, чистоту и когерентность тем при незначительной потере правдоподобия [151]. 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 [17, 172, 64], хотя никакие критерии качества поиска непосредственно не оптимизировались.

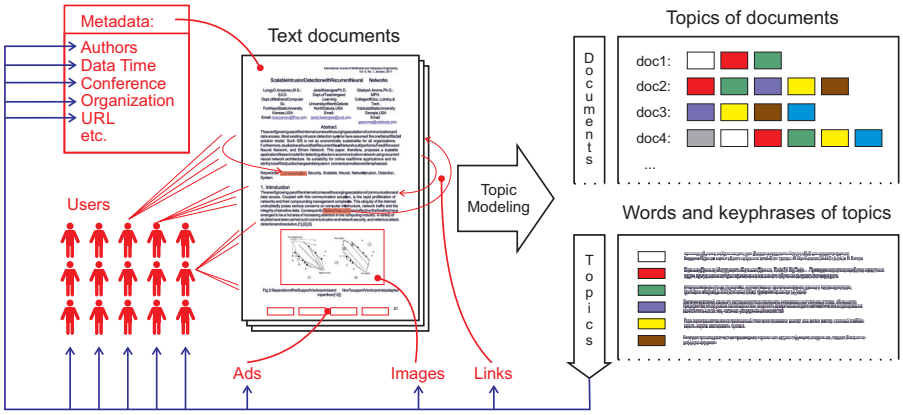


Рис. 10: Обычная тематическая модель определяет распределение тем в каждом документе $p(t|d)$ и распределение термов в каждой теме $p(w|t)$. Мультимодальная модель распространяет семантику тем на элементы всех остальных модальностей, в том числе нетекстовые.

8 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 [158, 167], теги [73], именованные сущности [105] — это примеры текстовых модальностей. В мультиязычных тематических моделях параллельных текстов модальностями являются языки [153]. Для анализа коротких текстов с опечатками используют модальность буквенных n-грамм, что позволяет улучшать качество информационного поиска [63]. Примерами нетекстовых модальностей являются (рис. 10): авторы [128], моменты времени [144, 179, 145], классы, жанры или категории [129, 182], цитируемые или цитирующие документы [48] или авторы [69], пользователи электронных библиотек, социальных сетей или рекомендательных систем [77, 136, 160, 175, 176], графические элементы изображений [32, 62, 82], рекламные объявления на веб-страницах [117].

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):

$$p(w\,|\,d) = \sum_{t\in T} p(w\,|\,t)\,p(t\,|\,d) = \sum_{t\in T} \varphi_{wt}\theta_{td}, \qquad w \in W_m, \quad d \in D. \tag{33}$$

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:

$$\sum_{m\in M}\tau_m\sum_{d\in D}\sum_{w\in W_m} n_{dw}\ln\sum_{t\in T}\varphi_{wt}\theta_{td} + R(\Phi,\Theta) \;\to\; \max_{\Phi,\Theta}; \tag{34}$$

$$\sum_{w\in W_m}\varphi_{wt}=1; \quad \varphi_{wt}\geqslant 0; \qquad \sum_{t\in T}\theta_{td}=1; \quad \theta_{td}\geqslant 0. \tag{35}$$

Теорема 8.1. *Пусть функция $R(\Phi,\Theta)$ непрерывно дифференцируема. Точка (Φ,Θ) локального экстремума задачи (34)–(35) удовлетворяет системе уравнений со вспомогательными переменными $p_{tdw}=p(t\,|\,d,w)$, если из решения исключить нулевые столбцы матриц Φ_m, Θ:*

$$p_{tdw} = \underset{t\in T}{\operatorname{norm}}\big(\varphi_{wt}\theta_{td}\big); \tag{36}$$

$$\varphi_{wt} = \underset{w\in W_m}{\operatorname{norm}}\left(n_{wt}+\varphi_{wt}\frac{\partial R}{\partial\varphi_{wt}}\right); \qquad n_{wt}=\sum_{d\in D}\tau_{m(w)}n_{dw}p_{tdw}; \tag{37}$$

$$\theta_{td} = \underset{t\in T}{\operatorname{norm}}\left(n_{td}+\theta_{td}\frac{\partial R}{\partial\theta_{td}}\right); \qquad n_{td}=\sum_{m\in M}\sum_{w\in W^m}\tau_m n_{dw}p_{tdw}. \tag{38}$$

Доказательство аналогично теореме 17.2, которая является частным случаем теоремы 8.1 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 [21].

Мультиязычные модели. Мультиязычные текстовые коллекции используются для кросс-язычного информационного поиска, когда по запросу на одном языке требуется найти семантически близкие документы на другом языке. Для связывания языков используются параллельные тексты или двуязычные словари. Первые мультиязычные тематические модели появились почти одновременно [46, 99, 110] 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 двух 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 [153].

На рис. 11 показаны некоторые из 400 тем, построенных по 216 175 парам русских и английских статей Википедии [149]. 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 [9] был предложен регуляризатор сглаживания. Он формализует предположение, что если слово u в языке k является переводом слова w из языка ℓ, то их распределения тем $p(t\,|\,u)$ и $p(t\,|\,w)$ должны быть

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

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:

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

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:

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

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

Связывание параллельных текстов сильнее улучшает качество поиска, чем оба способа учёта словарей [9]. 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. 12.

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

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

$$p(w\,|\,d) = \sum_{t\in T} p(w\,|\,t) \sum_{c\in C_d} p(t\,|\,c)p(c\,|\,d) = \sum_{t\in T}\sum_{c\in C_d} \varphi_{wt}\psi_{tc}\pi_{cd}. \tag{39}$$

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 полностью определяется тематикой документа и не зависит от самого документа и его категорий.

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

Задача максимизации регуляризованного правдоподобия:

$$\sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{t\in T}\sum_{c\in C_d} \varphi_{wt}\psi_{tc}\pi_{cd} + R(\Phi,\Psi,\Pi) \;\to\; \max_{\Phi,\Psi,\Pi}; \tag{40}$$

$$\sum_{w\in W} \varphi_{wt} = 1,\ \varphi_{wt} \geqslant 0; \quad \sum_{t\in T} \psi_{tc} = 1,\ \psi_{tc} \geqslant 0; \quad \sum_{c\in C_d} \pi_{cd} = 1,\ \pi_{cd} \geqslant 0. \tag{41}$$

Теорема 8.2. *Пусть функция $R(\Phi, \Psi, \Pi)$ непрерывно дифференцируема. Точка локального экстремума (Φ, Ψ, Π) задачи (40), (41) 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 17.2.

Модель трёхматричного разложения наиболее известна как *автор-тематическая модель* ATM (author-topic model), в которой порождающей модальностью являются авторы документов [128]. В *тематической модели тегирования документов* TWTM (tag weighted topic model) порождающей модальностью являются теги документа [80]. Аналогичная модель использовалась для обработки видеопотоков в [62]: 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 (39) 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 [129] и Labeled-LDA [124]. 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`.

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) [166] 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, имеющие спорадические всплески, 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:

$$R_1(\Phi \text{ или } \Theta) = -\tau_1 \sum_{i\in I} \sum_{t\in T} \ln p(t\,|\,i).$$

Во-вторых, предположим, что распределения $p(i\,|\,t)$ как функции времени меняются не слишком быстро и введём регуляризатор сглаживания:

$$R_2(\Phi \text{ или } \Theta) = -\tau_2 \sum_{i\in I} \sum_{t\in T} \big|p(i\,|\,t) - p(i{-}1\,|\,t)\big|.$$

Оба регуляризатора можно записать и как функцию от Φ, и как функцию от Θ. В случае регуляризатора $R_2(\Phi)$ формула M-шага имеет вид[2]

$$\varphi_{it} = \underset{i\in I}{\operatorname{norm}}\big(n_{it} + \tau_2\varphi_{it}\operatorname{sign}(\varphi_{i-1,t} - \varphi_{it}) + \tau_2\varphi_{it}\operatorname{sign}(\varphi_{i+1,t} - \varphi_{it})\big), \tag{42}$$

где функция sign возвращает $+1$ для положительного аргумента и -1 для отрицательного. Регуляризатор сглаживает значения в каждой точке временного ряда $p(i\,|\,t)$ по отношению к соседним точкам слева и справа.

9 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

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

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 взаимодействие объектов не сводится к парным взаимодействиям. В других случаях сложные взаимодействия всё же распадаются 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), которые 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 $p(t \mid v)$. 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 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. 13 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. Анало-

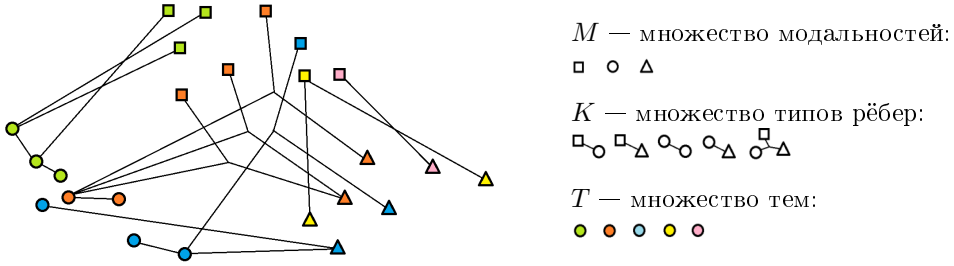


Рис. 13: Пример гиперграфа с вершинами трёх модальностей, рёбрами-транзакциями пяти типов и пятью темами.

гично 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. Более того, механизм регуляризации позволяет связать эти распределения и сделать их похожими.

Наконец, введём гипотезу условной независимости вершин в рёбрах (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 (34).

Гиперграфовая модель является широким обобщением обычных тематических моделей. В частности, она соответствует модели PLSA в случае, когда модальности 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 всегда 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 $R(\Phi, \Theta)$:

$$\sum_{k\in K} \tau_k \sum_{dx\in E_k} n_{kdx} \ln\Bigg(\sum_{t\in T} \theta_{td} \prod_{v\in x} \varphi_{vt}\Bigg) + R(\Phi,\Theta) \to \max_{\Phi,\Theta}; \tag{43}$$

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

Теорема 9.1. *Пусть функция $R(\Phi, \Theta)$ непрерывно дифференцируема. Точка локального максимума (Φ, Θ) задачи (43) 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:*

$$p_{tdx} = \underset{t\in T}{\operatorname{norm}}\Bigg(\theta_{td} \prod_{v\in x} \varphi_{vt}\Bigg). \tag{44}$$

$$\varphi_{vt} = \underset{v\in V_m}{\operatorname{norm}}\Bigg(n_{vt} + \varphi_{vt}\frac{\partial R}{\partial \varphi_{vt}}\Bigg); \qquad n_{vt} = \sum_{k\in K}\sum_{dx\in E_k} [v \in x]\, \tau_k n_{kdx} p_{tdx}; \tag{45}$$

$$\theta_{td} = \underset{t\in T}{\operatorname{norm}}\Bigg(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\Bigg); \qquad n_{td} = \sum_{k\in K}\sum_{dx\in E_k} \tau_k n_{kdx} p_{tdx}. \tag{46}$$

Доказательство. Воспользуемся леммой 3.2 о максимизации на единичных симплексах, выделив вспомогательные переменные p_{tdx}, определённые в (44):

$$\varphi_{vt} = \underset{v\in V_m}{\operatorname{norm}}\Bigg(\varphi_{vt}\sum_{k\in K}\tau_k \sum_{dx\in E_k} n_{kdx}\frac{\theta_{td}}{p(x\,|\,d)}\frac{\partial}{\partial \varphi_{vt}}\prod_{u\in x}\varphi_{ut} + \varphi_{vt}\frac{\partial R}{\partial \varphi_{vt}}\Bigg) =$$

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

$$\theta_{td} = \underset{t\in T}{\operatorname{norm}}\Bigg(\theta_{td}\sum_{k\in K}\tau_k \sum_{x\in d} n_{kdx}\frac{1}{p(x\,|\,d)}\prod_{v\in x}\varphi_{vt} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\Bigg) =$$

$$= \underset{t\in T}{\operatorname{norm}}\Bigg(\sum_{k\in K}\sum_{x\in d} \tau_k n_{kdx} p_{tdx} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\Bigg).$$

Теорема доказана.

Коэффициенты влияния. На практике возможны ситуации, когда определённые типы транзакций играют вспомогательную роль, и не должны влиять либо на тематику контейнеров, либо на распределения термов в темах. Например, текстовый комментарий в торговой транзакции не должен влиять на тематику продавца и покупателя, но при этом важно получать распределения слов комментариев в темах. В мультиязычной модели только модальность главного языка должна влиять на тематику мультиязычного документа. Музыкальные треки должны влиять на темы, если они находится в плей-листах пользователей, но не в альбомах исполнителей.

Для решения таких задач в формулы M-шага (45)–(46) вводятся *коэффициенты влияния* σ_{km}, σ_k, принимающие значения из отрезка $[0,1]$:

$$n_{vt} = \sum_{k \in K} \sum_{dx \in E_k} [v \in x]\, \tau_k \sigma_{k\mu(v)} n_{kdx} p_{tdx};$$
$$n_{td} = \sum_{k \in K} \sum_{dx \in E_k} \tau_k \sigma_k n_{kdx} p_{tdx}.$$

10 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 реализуется особенно просто, если классы считать модальностью. Регрессия на текстах показывает пример интересного приёма, когда дополнительные параметры модели (в данном случае коэффициенты линейной модели регрессии) пересчитываются итерационно после каждого прохода коллекции. Очень похожий приём используется и в модели CTM, которая выявляет попарные взаимосвязи между темами. Техника числовых модальностей используется в том случае, когда с текстом связана не одна числовая величина, которую необходимо предсказать, а числовая последовательность, которая моделируется смесью непрерывных вероятностных распределений.

Классификация. *Тематическая модель классификации* Dependency LDA [129] является байесовским аналогом модели (33) с модальностями термов W и классов C. Имеется обучающая выборка документов d, для каждого из которых известно подмножество классов $C_d \subset C$. Требуется классифицировать новые документы с неизвестным C_d. Для этого будем использовать *линейную вероятностную модель классификации*, в которой объектами являются документы d, признаки соответствуют темам t и принимают значения $\theta_{td} = p(t|d)$:

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

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. Признаковое описание нового документа θ_d вычисляется тематической моделью только по его термам.

Эксперименты в [129] показали, что тематические модели превосходят обычные методы многоклассовой классификации на больших текстовых коллекциях с большим числом несбалансированных, пересекающихся, взаимозависимых классов. В [148] те же выводы на тех же коллекциях были воспроизведены для мультимодальной ARTM. *Несбалансированность* означает, что классы могут содержать как малое, так и очень большое число документов. В случае *пересекающихся* классов документ может относиться 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 (33), 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 [88, 129]. 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 (16), получим формулы M-шага:

$$\varphi_{ct} = \underset{w\in W}{\operatorname{norm}}\left(n_{ct} + \tau |D_c| \frac{n_t \varphi_{ct}}{\sum_s n_s \varphi_{cs}}\right). \tag{47}$$

Частотная регуляризация использовалась в тематической модели Prior-LDA, которая была предложена в [129] 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 [92, 138].

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 (17) и приравнивая нулю его производную по v, получим формулы M-шага:

$$\theta_{td} = \operatorname*{norm}_t \Big(n_{td} + \tau v_t \theta_{td} \Big(y_d - \sum_{s \in T} v_s \theta_{sd} \Big) \Big);$$
$$v = (\Theta \Theta^T)^{-1} \Theta y.$$

Здесь 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 [138] 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. Темы поворачиваются 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 [28]. 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 (17), формула M-шага для θ_{td} принимает вид

$$\theta_{td} = \underset{t \in T}{\operatorname{norm}} \Big(n_{td} - \tau \sum_{s \in T} \Sigma^{-1}_{ts} \big(\ln \theta_{sd} - \mu_s \big) \Big), \tag{48}$$

где Σ^{-1}_{ts} — элементы обратной ковариационной матрицы. Параметры Σ, μ нормального распределения обновляются после каждого прохода коллекции, либо после каждого пакета документов 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, трудоёмкая операция обращения ковариационной матрицы выполняется относительно редко. В [28] использовалась LASSO-регрессия, чтобы получать разреженную ковариационную матрицу.

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 токеном.

Одним из примеров является задача тематического моделирования банковских транзакционных данных. В роли документов здесь 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. Благодаря названиям появляется возможность интерпретировать каждую тему с помощью двух списков — товаров, которые компании покупают для осуществления данной деятельности, и товаров, которые они продают как результат своей деятельности[3].

Кроме естественных дискретных модальностей контрагентов и названий, в данной задаче имеются ещё и числовые данные об объёмах сделок, которые могут нести важную информацию о деятельности компаний. 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

[3] Хрыльченко К. Я. Обобщенные модальности в вероятностных тематических моделях для транзакционных данных. Магистерская диссертация, ВМК МГУ, 2020.
`http://www.machinelearning.ru/wiki/images/5/50/Khrylchenko20msc.pdf`

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 (34) с произвольным непрерывно дифференцируемым регуляризатором $R(\Phi,\Theta)$. Добавим регуляризатор числовой модальности, определив его как логарифм правдоподобия выборок $\{y_{dmi}\}$:

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

Тогда система уравнений в теореме 8.1 преобразуется следующим образом: формулы E-шага (36) и M-шага по Φ (37) остаются в силе; к ним добавляются формулы E-шага и M-шага для числовых модальностей и изменяется формула M-шага по Θ:

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

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

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

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 разделе 9. 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

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

Соответственно модифицируется задача (43) максимизации log-правдоподобия транзакционных данных $X = \big\{(d_i,m_i,x_i,y_i)\colon i=1,\ldots,n\big\}$ для построения гипергра-

фовой тематической модели:

$$\sum_{(d,m,x,y)\in X} \ln\left(\sum_{t\in T} \theta_{td}\, p(y\,|\,t;\gamma_{tm}) \prod_{v\in x} \varphi_{vt}\right) + R(\Phi,\Theta) \to \max_{\Phi,\Theta,\Gamma}; \tag{49}$$

Теорема 10.1. *Пусть функция $R(\Phi,\Theta)$ непрерывно дифференцируема. Точка локального максимума (Φ,Θ,Γ) задачи (49) 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 ЕМ-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 Е-шаге.

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

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

11 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, гиперссылки, цитирование, совместное упоминание, комментирование, общие авторы, общие источники, близкие геолокации, и так далее. Предположение о сходстве тематики в парах документов легко формализуется с помощью регуляризатора матрицы Θ. Это может создавать технические трудности при обработке больших коллекций. В таких случаях ссылки можно переводить в модальность и переходить к регуляризатору 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 предположе-

ние 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 [48] 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 (17), имеет вид

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

Это ещё одна разновидность сглаживания. Вероятности θ_{td} в ходе итераций приближаются к вероятностям θ_{tc} документов, связанных с d.

Регуляризатор матрицы Θ становится 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: если мощность этой модальности окажется равной числу документов, то матрица Φ может не поместиться в оперативную память. Можно сократить эту модальность, оставив только наиболее влиятельные документы c, число ссылок на которые $n_c = \sum_d n_{dc}$ превышает выбранный порог.

Данная идея пришла из модели влияния научных публикаций LDA-post [48]. 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, чисто формальными или «данью вежливости». 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 случае используется регуляризатор. ARTM позволяет совмещать в модели 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 [177] определяются регионы популярности национальной кухни по постам пользователей Flickr. Другая иллюстрация из [93] показывает, что тематическая модель, учитывающая, из какого штата США пришло сообщение, точнее прослеживает путь урагана «Катрина».

Квадратичный регуляризатор матрицы Θ, предложенный в [177], 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 [93] 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:

$$R(\Phi) = -\frac{\tau}{2}\sum_{(u,v)\in E} w_{uv}\sum_{t\in T} n_t^2\Big(\frac{\varphi_{vt}}{|D_v|} - \frac{\varphi_{ut}}{|D_u|}\Big)^2. \tag{50}$$

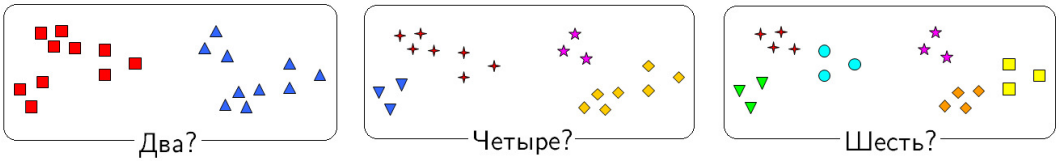


Рис. 14: Выбор числа кластеров в задачах кластеризации принципиально неоднозначен, как и выбор числа тем в задачах тематического моделирования.

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 [140] 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[4], регуляризация матрицы Φ приводит практически к тем же результатам, что и регуляризация Θ для моделей NetPLSA и iTopicModels.

12 Иерархические модели и выбор числа тем

Существует ли в реальных текстовых коллекциях «истинное» или «оптимальное» число тем? С одной стороны, число тем, о которых могут писать люди, ограничено. С другой стороны, в каждой теме имеется конечное число аспектов, и в каждом тексте авторы затрагивают лишь часть из них. Часто встречающиеся сочетания аспектов можно рассматривать и как отдельные темы, и как проявление общей темы. Кроме того, внутри любой темы могут существовать размытые кластерные структуры, для более чёткого определения которых может просто не хватать данных. Аналогичная проблема с неоднозначным выбором числа кластеров возникает и в задачах кластеризации, рис. 14.

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.

Чем больше коллекция, тем более мелкие семантические различия между темами возможно уловить. Однако даже при применении, казалось бы, строгих, статистических критериев для выявления значимых различий между темами, возникает произвол при выборе уровня значимости. Выбор этого порогового значения также является эвристикой, и от него зависит итоговое число тем. Все известные методы, претендующие на объективность определения числа тем, так или иначе содержат внутри себя параметр, от которого это самое число тем зависит.

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 последовательно дробятся на подтемы, а уровень

[4]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`

детальности (или *гранулированности*) тематического представления выбирается исходя из потребностей прикладной задачи.

Определение числа тем по внешним критериям. Единственным беспроигрышным вариантом определения оптимального числа тем является использование внешнего критерия. Такой подход используется, когда конечной целью (или одной из целей) тематического моделирования является решение задачи классификации [129], информационного поиска [64] или сегментации [126] 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 методом анализа является построение графика зависимости внешнего критерия от числа тем.

Энтропийное разреживание для отбора тем предложено в [150] для удаления незначимых тем из тематической модели. Идея заключается в разреживании распределения $p(t)$, 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 (17):

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

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

$$\theta_{td} = \underset{t \in T}{\text{norm}}\Big(n_{td}\Big(1 - \tau \frac{n}{n_t |T|}\Big)\Big). \tag{51}$$

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) [142] или процесса китайского ресторана (Chinese restaurant process, CRP) [31].

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 [152] 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 [152] также выяснилось, что регуляризатор отбора тем имеет полезный сопутствующий эффект: он удаляет из модели дублирующие, расщеплённые и линейно зависимые темы.

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 [30, 81, 98, 178, 122, 161, 162, 163, 164], оптимизация размера и структуры иерархии остаётся открытой проблемой; более того, оценивание качества иерархий — также открытая проблема [178].

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 [41] 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 коллекция подтемы могут иметь по несколько родительских тем, 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, $|S| > |T|$. Потребуем, чтобы родительские темы t хорошо приближались вероятностными смесями дочерних тем s:

$$\sum_{t\in T} n_t \operatorname{KL}_w\Big(p(w\,|\,t) \,\Big\|\, \sum_{s\in S} p(w\,|\,s)\,p(s\,|\,t)\Big) = \sum_{t\in T} n_t \operatorname{KL}_w\Big(\tfrac{n_{wt}}{n_t} \,\Big\|\, \sum_{s\in S} \varphi_{ws}\psi_{st}\Big) \;\to\; \min_{\Phi,\Psi},$$

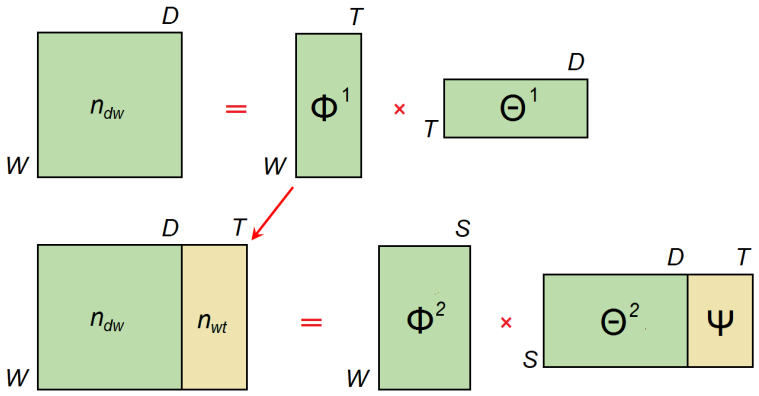


Рис. 15: Добавление второго уровня иерархии с множеством подтем S реализуется путём добавления в исходную коллекцию $|T|$ псевдодокументов с частотами термов n_{wt}. Матрица связей тем с подтемами $\Psi = \big(p(s|t)\big)$ образуется в столбцах матрицы Θ, соответствующих псевдодокументам.

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

Это задача матричного разложения для матрицы родительского уровня $\Phi^\ell = \Phi\Psi$. Обычно матричные разложения используются, чтобы приблизить матрицу высокого ранга произведением матриц более низкого ранга. Однако в данном случае, наоборот, матрица Φ^ℓ низкого ранга $|T|$ приближается произведением матриц $\Phi\Psi$, в котором матрица Φ должна иметь полный ранг $|S|$, чтобы дочерняя модель описывала коллекцию точнее, чем родительская. Регуляризатор связывает тематические модели соседних уровней ℓ и $\ell+1$ так, чтобы родительские темы φ_t^ℓ аппроксимировались линейными комбинациями дочерних тем φ_s с коэффициентами ψ_{st}:

$$R(\Phi,\Psi) = \tau \sum_{t\in T}\sum_{w\in W} n_{wt} \ln \sum_{s\in S} \varphi_{ws}\psi_{st}. \tag{52}$$

Задача максимизации $R(\Phi,\Psi)$ с точностью до обозначений совпадает с основной задачей тематического моделирования (9), если считать родительские темы t *псевдодокументами* с частотами термов $n_{wt} = \tau n_t \varphi_{wt}$. Это означает, что вместо добавления слагаемого в формулы М-шага данный регуляризатор можно реализовать ещё проще. Построив родительский уровень, надо добавить в коллекцию $|T|$ псевдодокументов, задав им в качестве частот термов значения n_{wt}. Матрица Ψ получится в столбцах матрицы Θ, соответствующих псевдодокументам [41], как показано на рис. 15.

Данный подход к моделированию иерархий реализован в библиотеке `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 (17), из которой следует формула М-шага для модели дочернего уровня:

$$\psi_{st} = \underset{s \in S}{\operatorname{norm}}\Big(n_{st} + \tau\Big(p(t\,|\,s) - \frac{1}{|T|}\Big)\Big). \tag{53}$$

Согласно этой формуле, условные вероятности $p(t\,|\,s)$, меньшие $\frac{1}{|T|}$, становятся ещё меньше, и при достаточно большом τ обнуляются [41].

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.

13 Моделирование сочетаемости слов

Гипотеза «мешка слов» является одним из самых критикуемых постулатов тематического моделирования. Она полностью игнорирует фундаментальное свойство *сочетаемости слов* (word co-occurrence) в естественном языке. Сочетаемость бывает двух видов: *контактная* и *дистантная*.

Для описания контактной сочетаемости из текста выделяются всевозможные *n-граммы* — последовательности из n подряд идущих слов. Среди n-грамм представляют интерес коллокации и словосочетания. *Коллокация* — это n-грамма, встречающаяся в корпусе гораздо чаще, чем можно было бы ожидать при чисто случайном соединении данных слов. *Словосочетание* — это n-грамма, слова которой связаны по смыслу и грамматически, и обозначают единое понятие. В тематических моделях тематику таких n-грамм лучше определять, не разделяя их на отдельные слова.

Дистантная сочетаемость означает частое появление слов в одних и тех же контекстах, например, в одном предложении или в соседних предложениях, но не обязательно вплотную друг к другу.

Многие исследования направлены на создание тематических моделей, учитывающих порядок слов. Из них наиболее важными представляются три направления.

Первое направление связано с выделением информативных n-грамм — коллокаций, словосочетаний, терминов, именованных сущностей. Темы, построенные на n-граммах, намного лучше интерпретируются, чем построенные на униграммах (отдельных словах). Проблема в том, что число всех n-грамм катастрофически быстро растёт с ростом объёма коллекции.

Второе направление связано с основной гипотезой *дистрибутивной семантики*: смысл слова определяется тем, в окружении каких слов оно чаще всего употребляется. Появление программы `word2vec` [94] стимулировало развитие *векторных представлений слов* (word embedding). Они находят массу применений благодаря тому, что семантически близким словам соответствуют близкие векторы. Тематические модели также способны строить векторные представления слов, обладающие этим свойством, в то же время сохраняя свойства интерпретируемости и разреженности.

Третье направление связано с *тематической сегментацией* и гипотезой, что текст на естественном языке состоит из последовательности монотематичных сообщений. В частности, каждое предложение чаще всего относится только к одной теме. Задачи сегментации рассматриваются в разделе 14.

распознавание образов в биоинформатике		теория вычислительной сложности	
unigrams	**bigrams**	**unigrams**	**bigrams**
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. 16: Примеры тем униграммной модели и соответствующих им тем биграммной модели (по коллекции статей научной конференции ММРО — «Математические методы распознавания образов»).

Модели контактной сочетаемости. Использование n-грамм, коллокаций или словосочетаний заметно улучшает интерпретируемость тем, что демонстрируется практически в каждой публикации по n-граммным тематическим моделям, см. например [67].

Первая биграммная тематическая модель BTM (bigram topic model) [158] 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) [167] устраняет эти недостатки. Условное распределение слов описывается вероятностной смесью $p(w|v,t) = \xi_{vwt}\varphi^v_{wt} + (1-\xi_{vwt})\varphi_{wt}$, 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

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

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. 16.

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 [52]. 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, обладающих свойствами (а) и (б).

Методы, предложенные в последующих работах, SegPhrase [84] и AutoPhrase [131], 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 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) [171].

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 совместного появления 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 битермов 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 (16)–(17) к этому регуляризатору даёт формулы М-шага:

$$\varphi_{wt} = \underset{w \in W}{\operatorname{norm}}\Big(n_{wt} + \tau \sum_{u \in W} n_{uw} p_{tuw}\Big); \tag{54}$$

$$p_{tuw} = \underset{t \in T}{\operatorname{norm}}\big(n_t \varphi_{wt}\varphi_{ut}\big). \tag{55}$$

Эти формулы интерпретируются как добавление *псевдо-документов*. Каждому слову $u \in W$ ставится в соответствие псевдо-документ d_u, объединяющий все контексты слова u, то есть это мешок слов, встретившихся рядом со словом u по всей коллекции. Число вхождений слова w в псевдо-документ d_u равно τn_{uw}. Вспомогательные переменные $p_{tuw} = p(t\,|\,u, w)$ соответствуют формуле Е-шага для псевдо-документа d_u, если доопределить его тематику как $\theta_{tu} = \underset{t}{\operatorname{norm}}(n_t \varphi_{ut})$. Другими словами, в модели битермов столбцы матрицы Θ, соответствующие псевдо-документам, образуются путём перенормировки строк матрицы Φ по формуле Байеса.

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 тематических моделей дистантной парной сочетаемости слов WTM (word topic model) [39] и WNTM (word network topic model) [183]. Любопытно, что более ранняя публикация [39] осталась незамеченной (видимо, как небайесовская), и статья [183] 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 log-правдоподобие для модели $p(w\,|\,d_u)$ 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, $n_{uw} = n_{wu}$).

Основное отличие этих моделей от модели битермов в том, что здесь в явном виде строится матрица Θ для псевдо-коллекции, тогда как в модели битермов $\Theta = \operatorname{diag}(\pi_1, \ldots, \pi_t)\Phi^{\mathsf{T}}$ и количество параметров вдвое меньше. Как показали эксперименты, модель WNTM немного превосходит модель битермов и существенно превосходит обычные тематические модели [183] на коллекциях коротких текстов. На коллекциях длинных документов тематические модели парной сочетаемости слов не дают значимых преимуществ перед обычными тематическими моделями.

Когерентность. Тема называется *когерентной* (согласованной), если наиболее частые термы данной темы часто встречаются рядом в документах коллекции [107]. Сочетаемость термов может оцениваться по самой коллекции D [100], или по сторонней коллекции, например, по Википедии [104]. Средняя когерентность тем считается хорошей мерой интерпретируемости тематической модели [108].

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

Запишем формулу полной вероятности $p(w\,|\,t) = \sum_v C_{wv}\varphi_{vt}$ и заменим в ней условную вероятность φ_{vt} частотной оценкой: $\hat{p}(w\,|\,t) = \sum_v C_{wv}\frac{n_{vt}}{n_t}$. Введём регуляризатор, требующий, чтобы параметры φ_{wt} тематической модели были согласованы с оценками $\hat{p}(w\,|\,t)$ в смысле кросс-энтропии:

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

Формула М-шага, согласно (16), принимает вид

$$\varphi_{wt} = \underset{w\in W}{\operatorname{norm}}\Big(n_{wt} + \tau \sum_{v\in W\setminus w} C_{wv} n_{vt}\Big). \tag{56}$$

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 [100] 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 [104] 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)

$$\mathrm{PMI}(u,v) = \ln \frac{|D| N_{uv}}{N_u N_v}, \tag{57}$$

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 [59]. Слова, встречающиеся

в схожих контекстах, имеют схожую семантику и, соответственно, должны иметь близкие векторы. Для формализации этого принципа в [94, 95] 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:

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

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 log-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 тематические модели BTM и WNTM. Модели семейства `word2vec` и другие модели векторных представлений слов также являются матричными разложениями [79, 116, 85]. 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 [119] 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. 17 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 [44], специально разработанную для поиска семантически близких документов.

Кроме того, ARTM позволяет обобщить тематические модели дистрибутивной семантики для мультимодальных коллекций [119]. Используя данные о парной сочетаемости термов различных модальностей, возможно строить интерпретируемые тематические векторные представления для всех модальностей. В то же время, привлечение дополнительной информации о других модальностях повышает качество решения задачи близости слов.

Ассоциация	**Результат ARTM**	**Результат word2vec**
king – boy + girl	*queen*, princess, lord, prince	*queen*, princess, regnant, kings
moscow – russia + spain	*madrid*, barcelona, aires, buenos	*madrid*, barcelona, valladolid, malaga
india – russia + ruble	*rupee*, birbhum, pradesh, madhaya	*rupee*, rupiah, devalued, debased
better – good + bad	really, something, thing, nothing	*worse*, easier, prettier, funnier
cars – car + computer	*computers*, software, servers, implementations	*computers*, software, hardware, microcomputers

Рис. 17: Примеры решения задач ассоциаций слов для моделей ARTM и word2vec. Приводятся четыре наиболее близкие ассоциации.

14 Моделирование связного 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 тематическая модель в целом покажется не настолько хорошо интерпретируемой, как ранжированные списки наиболее частотных слов в темах.

Тематические модели сегментации основаны на более реалистичных гипотезах о связности текста. Каждое предложение, как правило, относится к одной–двум темам. Следующее предложение, как правило, продолжает тематику предыдущего. Смена темы редко происходит между предложениями, часто между абзацами, ещё чаще — между секциями документа. Каждое предложение можно считать «мешком слов» или «мешком термов». Документ можно считать «мешком предложений».

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:

$$\sum_{d\in D} \sum_{s\in S_d} \ln \sum_{t\in T} \theta_{td} \prod_{w\in s} \varphi_{wt}^{n_{sw}} + R(\Phi,\Theta) \to \max_{\Phi,\Theta}, \tag{58}$$

при обычных ограничениях (10). В частном случае, когда каждый сегмент состоит только из одного слова, данная задача переходит в (11). Заметим также, что задача (58) является частным случаем построения тематической модели гиперграфа (43), в котором вершины являются словами, а рёбра — предложениями.

Теорема 14.1. *Пусть функция $R(\Phi,\Theta)$ непрерывно дифференцируема. Точка (Φ,Θ) локального экстремума задачи (58), (10) 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 [25] и для модели коротких сообщений Twitter-LDA [180]. 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. Как уже отмечалось выше, тематическая модель гиперграфа (43) может быть непосредственно применена для построения тематической модели предложений. На самом деле, ребром гиперграфа можно описать не только предложение, но и любое подмножество термов, 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 описывать рёбрами гиперграфа.

Можно использовать внешние лингвистические ресурсы — *тезаурусы* или *онтологии*, такие как WordNet, РуТез, Вики-Словарь и другие. 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 модель сегментации. Теперь рассмотрим более сложный случай, когда текст состоит из предложений, и требуется объединить их в более крупные тематические сегменты, границы которых заранее не определены.

Метод TopicTiling [126] основан на пост-обработке распределений $p(t\,|\,d, w_i)$, $i = 1, \ldots, n$, получаемых какой-либо тематической моделью, например, LDA. Определим тематику предложения s как среднюю тематику $p(t\,|\,d, w)$ всех его слов[6]. Посчи-

[6] Точнее, в [126] предлагалось для каждого слова выбирать наиболее вероятную тему. Оба варианта имеют право на существование. Какой из них лучше, пока не исследовано.

таем косинусную близость тематики для всех пар соседних предложений. 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 [126].

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. Некоторые требования к тематической модели удобнее выражать не через параметры модели φ_{wt} и θ_{td}, а через распределения $p_{tdw} = p(t \,|\, d, w)$, которые вычисляются на E-шаге. Например, это могут быть требования сходства тематики термов внутри предложений или между соседними предложениями. Они позволяют учитывать порядок слов в документах в обход гипотезы «мешка слов». В общем случае их удобно выражать с помощью регуляризатора $R(\Pi, \Phi, \Theta)$, где $\Pi = (p_{tdw})_{T \times D \times W}$ — трёхмерная матрица вспомогательных переменных.

Будем предполагать, что R является достаточно гладкой функцией 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 (15), матрица Π является функцией от Φ и Θ. Поэтому к регуляризатору $\tilde{R}(\Phi, \Theta) = R(\Pi(\Phi, \Theta), \Phi, \Theta)$ применима теорема 17.2. Однако систему уравнений удобно записывать через частные производные регуляризатора R, а не $\tilde{R}$.

Рассмотрим задачу максимизации регуляризованного log-правдоподобия

$$\sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \varphi_{wt} \theta_{td} + R(\Pi(\Phi, \Theta), \Phi, \Theta) \;\to\; \max_{\Phi, \Theta}, \tag{59}$$

при ограничениях неотрицательности и нормировки (10).

Теорема 14.2. *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 (59), (10) удовлетворяет системе уравнений со вспомогательными переменными p_{tdw} и $\tilde{p}_{tdw}$, если из решения исключить нулевые столбцы матриц Φ, Θ:*

$$p_{tdw} = \underset{t \in T}{\operatorname{norm}}\big(\varphi_{wt} \theta_{td}\big); \tag{60}$$

$$\tilde{p}_{tdw} = p_{tdw} \left(1 + \frac{1}{n_{dw}} \left(\frac{\partial R}{\partial p_{tdw}} - \sum_{z \in T} p_{zdw} \frac{\partial R}{\partial p_{zdw}}\right)\right); \tag{61}$$

$$\varphi_{wt} = \underset{w \in W}{\operatorname{norm}} \left(\sum_{d \in D} n_{dw} \tilde{p}_{tdw} + \varphi_{wt} \frac{\partial R}{\partial \varphi_{wt}}\right); \tag{62}$$

$$\theta_{td} = \underset{t \in T}{\operatorname{norm}} \left(\sum_{w \in d} n_{dw} \tilde{p}_{tdw} + \theta_{td} \frac{\partial R}{\partial \theta_{td}}\right). \tag{63}$$

Доказательство теоремы удобно разбить на три леммы.

Лемма 14.3. *Для функции* $p_{zdw}(\Phi,\Theta)=\frac{\varphi_{wz}\theta_{zd}}{\sum_t\varphi_{wt}\theta_{td}}$ *и любых* $t,z\in T$

$$\varphi_{wt}\frac{\partial p_{zdw}}{\partial\varphi_{wt}}=\theta_{td}\frac{\partial p_{zdw}}{\partial\theta_{td}}=p_{tdw}\big([z{=}t]-p_{zdw}\big)=\begin{cases}p_{tdw}(1-p_{tdw}), & t=z;\\ -p_{tdw}p_{zdw}, & t\neq z.\end{cases}$$

Доказательство. Воспользуемся формулами (15) для переменных p_{tdw}:

$$\begin{aligned}\varphi_{wt}\frac{\partial p_{zdw}}{\partial\varphi_{wt}}&=\varphi_{wt}\frac{\partial}{\partial\varphi_{wt}}\left(\frac{\varphi_{wz}\theta_{zd}}{\sum_u\varphi_{wu}\theta_{ud}}\right)=\varphi_{wt}\frac{[z{=}t]\theta_{td}\sum_u\varphi_{wu}\theta_{ud}-\theta_{td}\varphi_{wz}\theta_{zd}}{(\sum_u\varphi_{wu}\theta_{ud})^2}=\\&=p_{tdw}[z{=}t]-p_{tdw}p_{zdw}=p_{tdw}([z{=}t]-p_{zdw});\\ \theta_{td}\frac{\partial p_{zdw}}{\partial\varphi_{td}}&=\theta_{td}\frac{\partial}{\partial\theta_{td}}\left(\frac{\varphi_{wz}\theta_{zd}}{\sum_u\varphi_{wu}\theta_{ud}}\right)=\theta_{td}\frac{[z{=}t]\varphi_{wt}\sum_u\varphi_{wu}\theta_{ud}-\varphi_{wt}\varphi_{wz}\theta_{zd}}{(\sum_u\varphi_{wu}\theta_{ud})^2}=\\&=p_{tdw}[z{=}t]-p_{tdw}p_{zdw}=p_{tdw}([z{=}t]-p_{zdw}).\end{aligned}$$

Лемма доказана.

Введём вспомогательную функцию Q от переменных Π,Φ,Θ:

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

Лемма 14.4. *Пусть функция* $R(\Pi,\Phi,\Theta)$ *не зависит от переменных* p_{tdw} *при* $w\notin d$. *Тогда частные производные функции* $\tilde R(\Phi,\Theta)$ *выражаются через* Q_{tdw}:

$$\varphi_{wt}\frac{\partial\tilde R}{\partial\varphi_{wt}}=\varphi_{wt}\frac{\partial R}{\partial\varphi_{wt}}+\sum_{d\in D}p_{tdw}Q_{tdw};\qquad\theta_{td}\frac{\partial\tilde R}{\partial\theta_{td}}=\theta_{td}\frac{\partial R}{\partial\theta_{td}}+\sum_{w\in d}p_{tdw}Q_{tdw}.$$

Доказательство. Воспользуемся формулой дифференцирования сложной функции и тем, что $\dfrac{\partial p_{zdw'}}{\partial\varphi_{wt}}=0$ при $w\neq w'$; $\dfrac{\partial p_{zd'w}}{\partial\theta_{td}}=0$ при $d\neq d'$; $\dfrac{\partial R}{\partial p_{tdw}}=0$ при $w\notin d$:

$$\frac{\partial\tilde R}{\partial\varphi_{wt}}=\frac{\partial R}{\partial\varphi_{wt}}+\sum_{(z,d,w')}\frac{\partial R}{\partial p_{zdw'}}\frac{\partial p_{zdw'}}{\partial\varphi_{wt}}=\frac{\partial R}{\partial\varphi_{wt}}+\sum_{d\in D}\sum_{z\in T}\frac{\partial R}{\partial p_{zdw}}\frac{\partial p_{zdw}}{\partial\varphi_{wt}};\tag{64}$$

$$\frac{\partial\tilde R}{\partial\theta_{td}}=\frac{\partial R}{\partial\theta_{td}}+\sum_{(z,d',w)}\frac{\partial R}{\partial p_{zd'w}}\frac{\partial p_{zd'w}}{\partial\theta_{td}}=\frac{\partial R}{\partial\theta_{td}}+\sum_{w\in d}\sum_{z\in T}\frac{\partial R}{\partial p_{zdw}}\frac{\partial p_{zdw}}{\partial\theta_{td}}.\tag{65}$$

В силу леммы 14.3 справедливо тождество

$$\begin{aligned}\varphi_{wt}\sum_{z\in T}\frac{\partial R}{\partial p_{zdw}}\frac{\partial p_{zdw}}{\partial\varphi_{wt}}&=\theta_{td}\sum_{z\in T}\frac{\partial R}{\partial p_{zdw}}\frac{\partial p_{zdw}}{\partial\theta_{td}}=\\&=\sum_{z\in T}\frac{\partial R}{\partial p_{zdw}}p_{tdw}\big([z{=}t]-p_{zdw}\big)=p_{tdw}\left(\frac{\partial R}{\partial p_{tdw}}-\sum_{z\in T}p_{zdw}\frac{\partial R}{\partial p_{zdw}}\right)=p_{tdw}Q_{tdw}.\end{aligned}$$

Подстановка полученного выражения в (64) и (65) завершает доказательство.

Лемма 14.5. *Решение* Φ, Θ *задачи* (59) *удовлетворяет следующей системе уравнений относительно переменных* φ_{wt}, θ_{td} *и вспомогательных переменных* p_{tdw}:

$$p_{tdw} = \underset{t\in T}{\operatorname{norm}}\big(\varphi_{wt}\theta_{td}\big); \tag{66}$$

$$\varphi_{wt} = \underset{w\in W}{\operatorname{norm}}\left(\sum_{d\in D} n_{dw}p_{tdw} + \sum_{d\in D} Q_{tdw}p_{tdw} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}}\right); \tag{67}$$

$$\theta_{td} = \underset{t\in T}{\operatorname{norm}}\left(\sum_{w\in d} n_{dw}p_{tdw} + \sum_{w\in d} Q_{tdw}p_{tdw} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right). \tag{68}$$

Эта лемма является непосредственным следствием теоремы 17.2 и леммы 14.4.

Осталось только заметить, что выделение вспомогательной переменной $\tilde{p}_{tdw}$ согласно (61) позволяет переписать уравнения (67)–(68) в требуемом виде (62)–(63).

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 (16)–(17) 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 (61):

$$\tilde{p}_{tdw} = p_{tdw} - \tau\big(\tfrac{1}{|T|} - p_{tdw}\big).$$

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 (67)–(68):

$$\varphi_{wt} = \underset{w\in W}{\operatorname{norm}}\left(n_{wt} - \tau n_w\left(\frac{1}{|T|} - \frac{n_{wt}}{n_w}\right)\right);$$

$$\theta_{td} = \underset{t\in T}{\operatorname{norm}}\left(n_{td} - \tau n_d\left(\frac{1}{|T|} - \frac{n_{td}}{n_d}\right)\right).$$

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:

$$R(\Pi) = -\tau \sum_{d \in D} \sum_{s \in S_d} \sum_{t \in T} \ln \sum_{w \in s} n_{sw} p_{tdw}. \tag{69}$$

Опуская рутинные выкладки, приведём результат подстановки (69) в (61):

$$\tilde{p}_{tdw} = p_{tdw}\left(1 - \frac{\tau}{n_{dw}} \sum_{s \in S_d} \frac{n_{sw}}{n_s}\left(\frac{1}{p_{tds}} - \sum_{z \in T} \frac{p_{zdw}}{p_{zds}}\right)\right).$$

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.

15 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 каждого типа вырожденности нужны свои 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 [129], точность и полнота информационного поиска [174, 19, 9, 17], число найденных хорошо интерпретируемых тем [22], качество сегментации текстов [126]. В [42] 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 (9), либо через log-правдоподобие (34) каждой модальности m:

$$\mathscr{P}_m(D;p) = \exp\left(-\frac{1}{n_m}\sum_{d\in D}\sum_{w\in W^m} n_{dw}\ln p(w\,|\,d)\right), \tag{70}$$

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 численных значений, а также её зависимость не только от качества модели, но и 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 обучающей выборки $\mathscr{P}_m(D;p_D)$ является оптимистично сме-

щённой (заниженной) характеристикой 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 $9:1$ [33].

В ранних экспериментах было показано, что LDA существенно превосходит PLSA по перплексии, откуда был сделан вывод, что LDA меньше переобучается [33]. Позже было показано, что на больших коллекциях перплексия моделей PLSA и LDA отличается незначительно [91, 170, 86].

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 [5, 120] 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 [6] 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 [6], 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 [37]. Примеры хорошо интерпретируемых тем показаны на рис. 11 и рис. 16. Свойство интерпретируемости важно в информационно-поисковых системах, использующих 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 [106] экспертам предлагалось непосредственно оценивать полезность тем по трёхбалльной шкале, рассматривая списки слов, ранжированные по убыванию $p(w|t)$. В *методе интрузий* [37] 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 [34, 71, 111] 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) [21, 22].

Когерентность. Экспертные подходы необходимы на стадии исследований, но они затрудняют автоматическое построение хороших тематических моделей. В серии работ [106, 107, 108, 100] 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 [107, 108]. Численной мерой когерентности темы t является *поточечная взаимная информация* (57), вычисляемая по k наиболее вероятным словам темы:

$$\mathscr{C}_t = \frac{2}{k(k-1)} \sum_{i=1}^{k-1} \sum_{j=i}^{k} \mathrm{PMI}(w_i, w_j), \tag{71}$$

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) [104], либо по той же коллекции, по которой строится модель [100].

Разреженность и лексические 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 [71] предлагается считать *существенными* лишь те условные вероятности, значения которых выше, чем у равномерного распределения:

$$W_t = \left\{ w \in W \;\middle|\; \varphi_{wt} > \tfrac{1}{|W|} \right\};$$
$$T_d = \left\{ t \in T \;\middle|\; \theta_{td} > \tfrac{1}{|T|} \right\}.$$

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. Таким образом, недостатком данного подхода можно считать игнорирование большинства слов.

В [150] предлагается определять *лексическое 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 $p(t|w) = \varphi_{wt}\frac{n_t}{n_w}$.

Независимо от того, каким образом определяется лексическое ядро, введём три показателя, характеризующих разреженность матрицы Φ:

$$\text{pur}_t = \sum_{w \in W_t} p(w|t) \text{ — } \textit{чистота темы} \text{ (чем выше, тем лучше);}$$
$$\text{con}_t = \tfrac{1}{|W_t|} \sum_{w \in W_t} p(t|w) \text{ — } \textit{контрастность темы} \text{ (чем выше, тем лучше);}$$
$$\text{ker}_t = |W_t| \text{ — } \textit{размер ядра} \text{ (ориентировочный оптимум } \tfrac{|W|}{|T|}).$$

Показатели размера ядра, чистоты и контрастности для модели в целом определяются как средние 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 лексическим ядром.

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 профессиональных текстах 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 послужить сигналом к очистке коллекции или внесению исправлений в модель.

Такие критерии, как разреженность, размер ядра или доля фоновой лексики могут использоваться для контроля избыточного разреживания модели.

Различность тем. Введём функцию расстояния между темами $\rho(t,t')$ как распределениями $p(w|t)$ и $p(w|t')$. Например, это может быть расстояние Хеллингера, косинусное расстояние или расстояние Жаккара между лексическими ядрами двух тем. Определим для каждой темы t расстояние до ближайшей к ней темы t':

$$\mathscr{R}_t = \min_{t' \in T \setminus t} \rho(t,t').$$

Если расстояние до ближайшей темы $\mathscr{R}_t$ мало, то темы t и t' дублируют друг друга, и, возможно, следует предпринять усилия, чтобы объединить эти темы.

Если расстояния $\mathscr{R}_t$ велики для всех тем, то это означает, что все темы попарно существенно различны. Такую модель будем называть *хорошо декоррелированной*.

16 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, равенства математического ожидания нулю, гауссовости, постоянства дисперсии, и т. д. Статистический анализ вероятностных тематических моделей не настолько хорошо разработан, хотя есть отдельные работы в этом направлении [157, 96, 53]. В данном разделе мы введём семейство средневзвешенных статистик, позволяющих измерять семантическую однородность и загрязнённость темы, согласованность документа с темой или терма с темой. Неожиданно обнаружится, что перплексия тоже принадлежит этому семейству статистик.

Гипотеза условной независимости допускает три эквивалентных представления:

$$p(w,d\,|\,t) = p(w\,|\,t)\,p(d\,|\,t); \tag{72}$$

$$p(w\,|\,d,t) = p(w\,|\,t); \tag{73}$$

$$p(d\,|\,w,t) = p(d\,|\,t). \tag{74}$$

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 (72), (73), (74) приводят к различным конструкциям статистических тестов с разными возможностями применения. Рассмотрим их подробнее.

1. Равенство (72) трансформируется в нулевую гипотезу о том, что для заданной темы t совместное распределение термов в документах порождается факторизованным распределением:

$$H_0: \quad \hat{p}(w,d\,|\,t) \sim p(w\,|\,t)\,p(d\,|\,t). \tag{75}$$

Для проверки данной гипотезы можно использовать статистику взаимной информации, предложенную в [96]:

$$S_t = \mathrm{KL}\big(\hat{p}(w,d\,|\,t) \;\big\|\; p(w\,|\,t)p(d\,|\,t)\big) = \sum_{d\in D}\sum_{w\in d} \hat{p}(w,d\,|\,t)\ln\frac{\hat{p}(w,d\,|\,t)}{p(w\,|\,t)p(d\,|\,t)}.$$

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 (15) и частотной оценкой условной вероятности $\hat{p}(w,d\,|\,t) = \frac{n_{tdw}}{n_t}$. Подставим полученные выражения в S_t:

$$\frac{\hat{p}(w,d\,|\,t)}{p(w\,|\,t)\,p(d\,|\,t)} = \frac{p(t\,|\,d,w)\,\hat{p}(w\,|\,d)}{p(w\,|\,t)\,p(t\,|\,d)} = \frac{p_{tdw}}{\varphi_{wt}\theta_{td}}\hat{p}(w\,|\,d) = \frac{\hat{p}(w\,|\,d)}{p(w\,|\,d)};$$

$$S_t = \sum_{d\in D}\sum_{w\in d}\frac{n_{tdw}}{n_t}\ln\frac{\hat{p}(w\,|\,d)}{p(w\,|\,d)} = \operatorname*{avg}_{d,w}\left(n_{tdw}, \ln\frac{\hat{p}(w\,|\,d)}{p(w\,|\,d)}\right), \tag{76}$$

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 $\ell(w,d)$ положительна, когда $p(w\,|\,d) < \hat{p}(w\,|\,d)$. Потеря тем выше, чем хуже тематическая модель предсказывает вероятность появления 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. Если в модели оказалось слишком много семантически неоднородных тем, то целесообразно построить другую модель, увеличив число тем. Темы с аномально большими значениями S_t могут в дальнейшем расщепляться на более мелкие темы в иерархических тематических моделях.

2. Равенство (73) трансформируется в нулевую гипотезу о том, что эмпирическое распределение термов темы t в документе d порождается общим для всех документов распределением термов:

$$H_0: \quad \hat{p}(w\,|\,d,t) \sim p(w\,|\,t). \tag{77}$$

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:

$$S_{td} = \mathrm{KL}\big(\hat{p}(w\,|\,d,t)\;\big\|\;p(w\,|\,t)\big) = \operatorname*{avg}_{w\in d}\left(n_{tdw}, \ln\frac{\hat{p}(w\,|\,d)}{p(w\,|\,d)}\right). \tag{78}$$

Статистика S_{td} является мерой несогласованности документа d с темой t.

Чем больше значение S_{td}, тем сильнее документ d отклоняется от семантически однородного кластера темы t. Статистика S_{td} может применяться для отбора релевантных документов при суммаризации или визуализации темы, удаления нетематизируемых или «грязных» документов, выбора числа итераций по документу, выделения документов для формирования начальных приближений новых тем.

3. Равенство (74) трансформируется в нулевую гипотезу о том, что эмпирическое распределение терма w по документам в теме t порождается общим для всех термов распределением темы по документам:

$$H_0: \quad \hat{p}(d\,|\,w,t) \sim p(d\,|\,t). \tag{79}$$

Для проверки данной гипотезы относительно фиксированного терма w введём статистику на основе KL-дивергенции. Снова опуская выкладки, запишем её через

средневзвешенное логарифмической функции потерь:

$$S_{wt} = \mathrm{KL}\big(\hat{p}(d\,|\,w,t) \;\big\|\; p(d\,|\,t)\big) = \operatorname*{avg}_{d\in D}\left(n_{tdw}, \ln\frac{\hat{p}(w\,|\,d)}{p(w\,|\,d)}\right). \tag{80}$$

Статистика S_{wt} является мерой несогласованности терма w с темой t.

Аномально высокие значения S_{wt} говорят о том, что терм относится к общеупотребительной лексике. Для выделения таких термов в фоновые темы можно усилить регуляризаторы декоррелирования и сглаживания [151]. Аномально низкие значения S_{wt} говорят о том, что терм w является обязательным в теме и входит в её семантическое ядро. Темы, содержащие большое число таких термов, могут быть образованы шаблонными фразами, часто повторяющимися в текстах коллекции [96].

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 [16].

Обобщённые средневзвешенные статистики. Обобщим введённые выше статистики S_t, S_{td}, S_{wt} на случай произвольной функции потерь $\ell(d,w)$:

$$\begin{aligned}
S_t &= \operatorname*{avg}_{d,w}\big(n_{tdw}, \ell(d,w)\big) \text{ — неоднородность темы } t \text{ в коллекции;}\\
S_{td} &= \operatorname*{avg}_{w\in d}\big(n_{tdw}, \ell(d,w)\big) \text{ — несогласованность документа } d \text{ с темой } t;\\
S_{wt} &= \operatorname*{avg}_{d\in D}\big(n_{tdw}, \ell(d,w)\big) \text{ — несогласованность терма } w \text{ с темой } t.
\end{aligned}$$

При логарифмической функции потерь $\ell(d,w) = \frac{\hat{p}(w\,|\,d)}{p(w\,|\,d)}$ обобщённые статистики S_t, S_{td}, S_{wt} переходят в (76), (78), (80) соответственно.

В общем случае от функции потерь потребуем, чтобы она не зависела от темы и чтобы значение $\ell(d,w)$ было тем выше, чем хуже тематическая 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 предположением. Некоторые языковые явления могут формально её нарушать, не приводя к семантической неоднородности. Например, явление *повторяемости слов* (word burstiness) — если слово встретилось в тексте один раз, то оно с большой вероятностью встретится ещё [49, 78]. Тема может содержать много синонимов, из которых каждый автор употребляет лишь некоторые. Тема может обладать несколькими аспектами, каждый со своей лексикой, однако в тексте может идти речь только о части аспектов. Вследствие этих явлений в документе может встретиться намного меньше слов темы, при этом некоторые из них будут встречаться намного чаще, чем можно было бы ожидать в условиях строгого выполнения гипотезы условной независимости.

Статистики S_t, S_{td}, S_{wt} нетрудно сделать толерантными к повторяемости слов, если заменить частоты слов в документах бинарными индикаторами: $n_{dw} := [n_{dw} \geqslant 1]$. Для этого достаточно вместо весов $n_{tdw} = n_{dw}p_{tdw}$ взять веса p_{tdw}:

$$S_t = \operatorname*{avg}_{d,w}\big(p_{tdw}, \ell(d,w)\big); \qquad S_{td} = \operatorname*{avg}_{w\in d}\big(p_{tdw}, \ell(d,w)\big); \qquad S_{wt} = \operatorname*{avg}_{d\in D}\big(p_{tdw}, \ell(d,w)\big).$$

Чтобы функцию потерь также сделать толерантной к повторяемости, будем сравнивать модельную вероятность $p(w\,|\,d)$ с частотной оценкой $\hat{p}(w\,|\,d) = \frac{n_{dw}}{n_d}$, в которой счётчик n_{dw} заменён единицей или, для общности, параметром α (уменьшение α делает статистики ещё более толерантными к нарушениям нулевой гипотезы):

$$\ell(d,w) = \big[p(w\,|\,d) < \tfrac{\alpha}{n_d}\big].$$

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 $\ell(d,w) = \frac{1}{\lambda(\lambda+1)}\big(\big(\frac{\hat{p}(w\,|\,d)}{p(w\,|\,d)}\big)^{\lambda} - 1\big)$ обобщённые средневзвешенные статистики переходят в дивергенцию Кресси–Рида. Это параметрическое семейство статистик используется для проверки гипотез о согласии эмпирического распределения с заданным дискретным распределением [43]. 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 функциями пары дискретных распределений, за исключением расстояния Хеллингера. Преимущество параметрического семейства в том, что свободой выбора параметра λ можно распорядиться для оптимизации какого-либо внешнего критерия качества.

17 Особенности реализации EM-алгоритма

EM-алгоритм в тематическом моделировании — это метод простых итераций для решения системы уравнений вида (15)–(17) в случае обычной двухматричной модели или (36)–(38) для мультимодальной модели или (44)–(46) для гиперграфовой модели. Реализации итерационного процесса могут отличаться порядком вычислений по формулам E-шага и M-шага. В рациональном варианте (Алгоритм 2, стр. 11) 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 данном разделе 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 4. 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 [60], позже для модели PLSA в [26]. 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 [60]

Алгоритм 4. 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 $p_{tdw} := \underset{t\in T}{\operatorname{norm}}\big(\varphi_{wt}\theta_{td}\big)$ для всех $w \in d,\ t \in T$;
7 $\theta_{td} := \underset{t\in T}{\operatorname{norm}}\big(\sum\limits_{w\in d} n_{dw}p_{tdw} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\big)$ для всех $t \in T$;
8 **пока** θ_d не сойдётся;
9 $\tilde{n}_{wt} := \tilde{n}_{wt} + n_{dw}p_{tdw}$ 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 $\varphi_{wt} := \underset{w\in W}{\operatorname{norm}}\big(n_{wt} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}}\big)$ для всех $w \in W,\ t \in T$;
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 $\varphi_{wt} := \underset{w\in W}{\operatorname{norm}}\big(n_{wt} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}}\big)$ для всех $w \in W,\ t \in T$;

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 [56]. Пока один процесс занят формированием нового приближения матрицы Φ, остальные процессы

продолжают использовать её предыдущее приближение для обработки пакетов и обновления счётчиков $\tilde{n}_{wt}$.

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.

Улучшение сходимости. Метод простой итерации в общем случае не гарантирует сходимость к стационарной точке регуляризованного правдоподобия [11]. Этот недостаток исправляется простой модификацией, не требующей дополнительных затрат времени или памяти. Вместо текущих значений φ_{wt} и θ_{td} в регуляризационных поправках М-шага надо подставлять их частотные оценки $\hat{\varphi}_{wt} = \frac{n_{wt}}{n_t}$ и $\hat{\theta}_{td} = \frac{n_{td}}{n_d}$ — те самые, которые вычисляются в модели PLSA:

$$\varphi_{wt} = \underset{w \in W}{\operatorname{norm}}\left(n_{wt} + \hat{\varphi}_{wt}\frac{\partial R(\hat{\Phi}, \hat{\Theta})}{\partial \varphi_{wt}}\right);$$
$$\theta_{td} = \underset{t \in T}{\operatorname{norm}}\left(n_{td} + \hat{\theta}_{td}\frac{\partial R(\hat{\Phi}, \hat{\Theta})}{\partial \theta_{td}}\right).$$

Эксперименты в [11] 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.

Исключение матрицы Θ из модели. Вычисление тематического вектора документа $\theta_d = (\theta_{td})_{t \in T}$ требует многих ЕМ-итераций по всем термам документа. Нельзя ли сделать так, чтобы распределение θ_d вычислялось за один линейный проход? Утвердительный ответ на этот вопрос дан в статье [10]. Требование, чтобы элементы матрицы Θ выражались в явном виде через элементы матрицы Φ, играет роль регуляризатора, хотя формально не является критерием вида $R(\Phi, \Theta) \to \max$.

Такой подход к тематическому моделированию имеет несколько преимуществ. Решение получается более устойчивым, поскольку в модель вводится «естественный регуляризатор». Вектор θ_d для любого документа при необходимости может быть получен быстро, за одно линейное прочтение документа. Фактически, матрица Θ полностью исключается из модели, что приводит к сокращению размерности модели и уменьшению переобучения. В моделях с двумя матрицами недостаточное качество матрицы Φ теоретически может быть скомпенсировано итерационным процессом подгонки каждого столбца θ_d матрицы Θ под конкретный документ d. Когда матрица Θ непосредственно зависит от матрицы Φ, такая подгонка становится невозможной. Кроме того, размер матрицы Θ линейно зависит от числа документов в коллекции, тогда как размер словаря увеличивается по сублинейному степенному закону Хипса [50]. Рост словаря может быть ограничен и принудительно, путём отбрасывания наименее частотных слов. Таким образом, не только размерность модели уменьшается, но и сокращается темп её роста при расширении коллекции.

Зависимость $\theta_{td}(\Phi)$ может быть получена, например, из требования, чтобы тематический вектор документа d совпадал с результатом первой итерации ЕМ-алгоритма

без регуляризации при равномерном начальном приближении $\theta_{td}^0 = \frac{1}{|T|}$:

$$\theta_{td}(\Phi) = \underset{t\in T}{\operatorname{norm}}\left(\sum_{w\in W} n_{dw}p_{tdw}\right) = \sum_{w\in d}\frac{n_{dw}}{n_d}\frac{\varphi_{wt}\theta_{td}^0}{\sum_s \varphi_{ws}\theta_{sd}^0} = \sum_{w\in d} p_{dw}\frac{\varphi_{wt}}{\sum_s \varphi_{ws}}, \tag{81}$$

где $p_{dw} = \frac{n_{dw}}{n_d} = \hat{p}(w\,|\,d)$ — частотная оценка условной вероятности терма в документе.

Рассмотрим сначала общий случай функциональной зависимости $\theta_{td}(\Phi)$.

Теорема 17.1. *Пусть функции $\theta_{td}(\Phi)$ и $R(\Phi,\Theta)$ непрерывно дифференцируемы. Тогда точка Φ локального экстремума задачи (11) с ограничениями (10) и дополнительными ограничениями-равенствами $\theta_{td} = \theta_{td}(\Phi)$ удовлетворяет системе уравнений со вспомогательными переменными $p_{tdw} = p(t\,|\,d,w)$, n_{td} и n_{wt}, если из решения исключить нулевые столбцы матриц Φ, Θ:*

$$p_{tdw} = \underset{t\in T}{\operatorname{norm}}\big(\varphi_{wt}\theta_{td}\big); \tag{82}$$

$$n_{td} = \sum_{w\in d} n_{dw}p_{tdw}; \tag{83}$$

$$n_{wt} = \sum_{d\in D} n_{dw}p_{tdw}; \tag{84}$$

$$\varphi_{wt} = \underset{w\in W}{\operatorname{norm}}\left(n_{wt} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}} + \varphi_{wt}\sum_{d\in D}\sum_{s\in T}\left(\frac{n_{sd}}{\theta_{sd}} + \frac{\partial R}{\partial \theta_{sd}}\right)\frac{\partial \theta_{sd}}{\partial \varphi_{wt}}\right). \tag{85}$$

Доказательство. Воспользуемся необходимым условием экстремума задачи (11) с ограничениями (10), которые даёт лемма 5.4. Формула E-шага (24) совпадает с (82). Рассмотрим функционал, максимизируемый на M-шаге (25):

$$Q(\Phi) = \sum_{d\in D}\sum_{u\in W}\sum_{s\in T} n_{du}p_{sdu}\big(\ln\varphi_{us} + \ln\theta_{sd}(\Phi)\big) + R(\Phi,\Theta(\Phi)) \;\to\; \max_{\Phi}.$$

Запишем частные производные Q по параметрам φ_{wt}, выделяя в формулах выражения n_{wt} и n_{sd} согласно (84) и (83) соответственно:

$$\varphi_{wt}\frac{\partial Q}{\partial \varphi_{wt}} = \sum_{d\in D} n_{dw}p_{tdw} + \sum_{d,s,u} n_{du}p_{sdu}\frac{\varphi_{wt}}{\theta_{sd}}\frac{\partial \theta_{sd}}{\partial \varphi_{wt}} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}} + \varphi_{wt}\sum_{d,s}\frac{\partial R}{\partial \theta_{sd}}\frac{\partial \theta_{sd}}{\partial \varphi_{wt}} =$$

$$= n_{wt} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}} + \varphi_{wt}\sum_{d,s}\left(\frac{n_{sd}}{\theta_{sd}} + \frac{\partial R}{\partial \theta_{sd}}\right)\frac{\partial \theta_{sd}}{\partial \varphi_{wt}}.$$

Отсюда и из леммы 3.2 о максимизации на единичных симплексах следует формула M-шага (85).

Теорема доказана.

Таким образом, модификация EM-алгоритма коснулись только формулы M-шага, причём аддитивная поправка к частотным оценкам условных вероятностей $p(w\,|\,t)$ имеет вид, аналогичный регуляризационным поправкам.

Теперь вернёмся к частному случаю (81). Чтобы применить к нему теорему 17.1, найдём частную производную:

$$\frac{\partial \theta_{sd}}{\partial \varphi_{wt}} = \frac{\partial}{\partial \varphi_{wt}}\left(\frac{p_{wd}\varphi_{ws}}{\sum_v \varphi_{wv}}\right) = p_{wd}\frac{\delta_{st}\sum_v \varphi_{wv} - \varphi_{ws}}{\big(\sum_v \varphi_{wv}\big)^2} = p_{wd}h_w\big(\delta_{st} - \varphi_{ws}h_w\big),$$

где $\delta_{st} = [s{=}t]$ — символ Кронекера, $h_w = \big(\sum_t \varphi_{wt}\big)^{-1}$. Подставим это выражение в (85) и перепишем уравнения в порядке, удобном для проведения вычислений методом простых итераций [10]:

$$
\begin{aligned}
h_w &= \big(\textstyle\sum_t \varphi_{wt}\big)^{-1};\\
\theta_{td} &= \sum_{w\in d} p_{dw}\varphi_{wt}h_w;\\
p_{tdw} &= \operatorname*{norm}_{t\in T}\big(\varphi_{wt}\theta_{td}\big);\\
c_{td} &= \frac{1}{\theta_{td}}\sum_{w\in d} n_{dw}p_{tdw} + \frac{\partial R}{\partial \theta_{td}};\\
\gamma_{dw} &= \sum_{t\in T} \varphi_{wt}c_{td};\\
p'_{tdw} &= p_{tdw} + n_d^{-1}\varphi_{wt}h_w(c_{td} - h_w\gamma_{dw});\\
\varphi_{wt} &= \operatorname*{norm}_{w\in W}\bigg(\sum_{d\in D} n_{dw}p'_{tdw} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}}\bigg).
\end{aligned}
$$

Вычисление переменных θ_{td}, p_{tdw}, c_{td}, γ_{dw} и p'_{tdw} образует E-шаг и занимает $O(n_d|T|)$ операций для каждого документа d, как и в обычном EM-алгоритме. Все эти переменные, относящиеся к конкретному документу d, можно удалять из памяти по окончании его обработки. Вычисление переменных φ_{wt} происходит по обычной формуле, за исключением того, что вместо условных вероятностей p_{tdw} подставляются переменные p'_{tdw}. Таким образом, модификация EM-алгоритма не приводит к существенному увеличению ни времени его работы, ни расхода памяти.

Обработка каждого документа делается за два прохода. На первом проходе вычисляются только переменные θ_{td}. На втором проходе вычисляются все остальные переменные, используемые в конечном итоге для вычисления p'_{tdw} и обновления φ_{wt}. Если требуется только найти тематический вектор документа θ_d, не обновляя матрицу Φ, то второй проход делать не нужно. Переменные h_w также можно не вычислять каждый раз, формируя их после очередного обновления матрицы Φ. Таким образом, рассмотренная модификация EM-алгоритма позволяет тематизировать новые документы максимально быстро.

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:

$$\sum_{d\in D}\sum_{w\in d} n_{dw}\ell\Big(\sum_{t\in T}\varphi_{wt}\theta_{td}\Big) + R(\Phi,\Theta) \;\to\; \max_{\Phi,\Theta}. \tag{86}$$

Теорема 17.2. *Решение Φ,Θ задачи* (86) *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*:*

$$p_{tdw} = \varphi_{wt}\theta_{td}\ell'\big(p(w|d)\big); \tag{87}$$

$$\varphi_{wt} = \underset{w\in W}{\operatorname{norm}}\bigg(\sum_{d\in D} n_{dw}p_{tdw} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}}\bigg); \tag{88}$$

$$\theta_{td} = \underset{t\in T}{\operatorname{norm}}\bigg(\sum_{w\in d} n_{dw}p_{tdw} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\bigg). \tag{89}$$

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 дивергенций Кресси–Рида.

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 [23] 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).

18 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.

```
# Import all necessary tools and data
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.datasets import fetch_20newsgroups
from numpy import array
import artm
# Extract data using sklearn and numpy
cv = CountVectorizer(max_features=1000, stop_words='english')
n_wd = array(cv.fit_transform(fetch_20newsgroups().data).todense()).T
vocabulary = cv.get_feature_names()
# Create batches and dictionary
bv = artm.BatchVectorizer(data_format='bow_n_wd',
                          n_wd=n_wd,
                          vocabulary=vocabulary)
# Learn simple PLSA model
model = artm.ARTM(num_topics=15, dictionary=bv.dictionary)
model.scores.add(artm.PerplexityScore(name='perp',
                                      dictionary=bv.dictionary))
model.fit_offline(bv, num_collection_passes=20)
# Print perplexity values by iterations
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 (30) с небольшим обобщением:

$$\varphi_{wt} = \underset{w \in W}{\text{norm}}\big(n_{wt} + \tau\beta_w f(\varphi_{wt})\big).$$

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 разреженности* 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 (31), 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 (32). От пользователя требуется указать коэффициент регуляризации τ и список модальностей, на которые нужно воздействовать.

`TopicSelectionThetaRegularizer` — *регуляризатор отбора тем в матрице* Θ, реализован по формуле (51). 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 (42). 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 (50). 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 (56) и в качестве параметра требует словарь парной сочетаемости слов C_{wv} (собрать его можно с помощью встроенного парсера).

`BitermsPhiRegularizer` — *регуляризатор битермов*, реализован по формулам (54)–(55) 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 (47). 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 (53).

`TopicSegmentationPtdwRegularizer` — *регуляризатор E-шага для разреживания сегментов в матрицах* (p_{tdw}), определяемый по формуле (69).

`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 4. 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 (70). 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 (71).

`TopicKernelScore` — *ядровые характеристики тем:* чистота pur_t, контрастность con_t, размер ядра ker_t оценивающие различность и, косвенно, интерпретируемость каждой темы t, см. стр. 75. 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 выше 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.

19 Разведочный информационный поиск

Важным приложением тематического моделирования является *информационный поиск* (information retrieval) [174, 19]. Современные поисковые системы предназначены, главным образом, для поиска конкретных ответов на короткие текстовые запросы. 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 должна быть удобно систематизированная информация, «дорожная карта» предметной области.

Для этих случаев подходит парадигма *разведочного информационного поиска* (exploratory search) [89, 168]. 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. Это требует затрат времени и высокой квалификации. При отсутствии инструмента для получения «общей картины» остаётся сомнение, что какие-то важные аспекты изучаемой проблемы

так и не были найдены. Если образно представить итеративный поиск как блуждание по лабиринту знаний, то разведочный поиск — это средство автоматического построения 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 [13]. 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 весьма скромной технике. Технологии информационного поиска на основе тематического моделирования в настоящее время находятся в стадии исследований и разработок [139, 28, 115, 35, 18, 160].

В литературе по разведочному поиску тематическое моделирование стали использовать относительно недавно [130, 58, 127, 147], а многие обзоры о нём вообще не упоминают [54, 123, 137, 68, 90, 65]. В статье [147] важными преимуществами тематических моделей называются гибкость, возможности визуализации и навигации. В то же время, в качестве недостатков отмечаются проблемы с интерпретируемостью тем, трудности с модификацией тематической модели при поступлении новых документов и высокая вычислительная сложность. Эти проблемы относятся к устаревшим методам и успешно решены в последние годы: десятки новых моделей разработаны для улучшения интерпретируемости; онлайновые алгоритмы способны обрабатывать большие коллекции и потоки документов за линейное время [97, 26, 148]. С другой стороны, в работах по тематическому моделированию разведочный поиск часто называют одним из важнейших приложений, а оценки качества поиска используют для валидации моделей [174, 19]. Однако эти исследования пока не привели к созданию общедоступных систем разведочного поиска. Всё это говорит о разобщённости научных сообществ, разрабатывающих эти два направления. Тенденция к их сближению наметилась лишь в последние годы.

Тематическая модель для разведочного поиска должна удовлетворять многим требованиям одновременно. Она должна состоять из хорошо интерпретируемых тем, поскольку темы интенсивно используются в пользовательском интерфейсе для навигации по коллекции и визуализации результатов поиска. Она должна быть разреженной, чтобы каждый документ состоял из небольшого числа тем — это необходимо для эффективного хранения инвертированного индекса. Она должна быть иерархической, чтобы пользователь мог получить представление о тематической структуре предметной области на любом уровне детализации. Она должна автоматически определять число тем на каждом уровне иерархии и автоматически создавать и именовать новые темы. Она должна быть мультиграммной, так как выделение ключевых фраз и терминов существенно улучшает интерпретируемость темы. Она должна быть мультиязычной в тех приложениях, где требуется кросс-язычный поиск, например, при анализе патентных баз. Она должна быть мультимодальной, чтобы учитывать метаданные документов, включая авторов и цитирование. Она должна быть тем-

поральной, чтобы выявлять динамику развития тем. Она должна быть сегментирующей, чтобы не только находить релевантные документы, но и указывать в них конкретные сегменты. Она должна быть обучаемой по оценкам асессоров или логам пользователей, чтобы постоянно улучшать качество поиска. Наконец, реализация должна быть онлайновой, параллельной и распределённой, чтобы эффективно обрабатывать большие коллекции текстов. Таким образом, многие возможности должны быть объединены для создания сервисов разведочного поиска.

Качество разведочного поиска. Модель ARTM для разведочного поиска была предложена в [17] и улучшена в [172, 64]. Для измерения качества разведочного тематического поиска использовались критерии точности и полноты на основе оценок асессоров. Для оценивания была составлена выборка запросов — 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 на коллекциях 175 тысяч статей русскоязычного коллективного блога `habrahabr.ru` и 760 тысяч статей англоязычного блога `techcrunch.com` показали, что тематический поиск находит больше релевантных документов, чем асессоры, сокращая среднее время поиска с получаса до секунды. Комбинирование регуляризаторов декоррелирования, разреживания и сглаживания вместе с модальностями n-грамм, авторов и категорий значимо улучшает качество поиска и позволяет достичь точности выше 80% и полноты выше 90%.

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 400 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 «дорожной карты», либо тематическая иерархия, либо динамика развития тем во времени, либо графовая структура взаимосвязей между темами, документами, авторами или иными модальностями, либо сегментная структура отдельных документов.

Графическая визуализация больших данных практически бесполезна в статичном исполнении, но может оказаться мощным когнитивным средством в случае интерактивной реализации. Это было понято более 20 лет назад и сформулировано Беном Шнейдерманом в виде *мантры визуального поиска информации*: «сначала крупный план, затем масшабирование и фильтрация, детали по требованию»[8] [135].

Графическое отображение результатов тематического моделирования и разведочного поиска согласуется с концепцией *дальнего чтения* (distant reading), предложенной социологом литературы Франко Моретти [102]. Он противопоставляет этот способ изучения текстов нашему обычному чтению (close reading). Невозможно про-

[7] `http://textvis.lnu.se` — интерактивный обзор средств визуализации текстов.

[8] Visual Information Seeking Mantra: «Overview first, zoom and filter, details on demand» [135].

читать миллионы книг или статей, но вполне возможно применить статистические методы и графическую визуализацию, чтобы понять в общих чертах, о чём вся эта литература, 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»[9].

Для библиотеки `BigARTM` в настоящее время развивается инструмент визуализации с веб-интерфейсом VisARTM[10], поддерживающий важнейшие формы представления тематических моделей. Интересной возможностью VisARTM является построение *спектра тем* — ранжированного списка тем, в котором семантически близкие темы находятся рядом. Группирование тем по смыслу помогает пользователям быстрее находить темы, акцентируя внимание на различиях между близкими темами.

20 Заключение

Тематическое моделирование является одним из инструментов статистического анализа текстов. За два десятилетия интенсивных исследований созданы сотни тематических моделей. Многие из них удалось включить в данный обзор.

Несмотря на успехи тематических моделей, о которых сообщают научные публикации, на практике используется в основном устаревшая модель латентного размещения Дирихле (LDA). Большое разнообразие моделей, сложность их математического описания на языке байесовского обучения, несовместимость реализаций создают барьеры для практического применения более широкого спектра моделей.

Теория *аддитивной регуляризации тематических моделей* (ARTM) и проект с открытым кодом `BigARTM` нацелены на преодоление этих барьеров. В ARTM тематическая модель определятся простой вероятностной порождающей моделью, описывающей структуру пространства параметров, и «мешком регуляризаторов», задающих дополнительные требования. Регуляризаторы аддитивны, взаимозаменяемы, их легко комбинировать и переносить из одних моделей в другие, что приводит к модульной технологии моделирования. Создание модели с требуемыми свойствами под конкретное приложение не требует ни трудоёмких математических выкладок, ни создания алгоритмически сложного кода.

В байесовском обучении дополнительные требования встраиваются в структуру порождающей модели, сильно усложняя вывод и не оставляя возможностей для модульной реализации. Байесовский вывод для большинства моделей, описанных в данном обзоре, требует нескольких страниц выкладок. Язык классической регуляризации оказывается не менее выразительным, но намного более простым. ARTM сокращает вывод буквально до нескольких строк. Единственное, что мы при этом теряем — возможность оценить не только значения параметров модели, но и их распределения. В практике тематического моделирования эта возможность никогда толком не используется, так что потеря не велика.

[9] «*Distant reading* is not an obstacle but a specific form of knowledge: fewer elements, hence a sharper sense of their overall interconnection. Shapes, relations, structures. Forms. Models.» [102].

[10] Федоряка Д. С. Технология интерактивной визуализации тематических моделей. Бакалаврская диссертация, МФТИ, 2017 (`www.MachineLearning.ru/wiki/images/d/d8/Fedoriaka17bsc.pdf`).

Доминирование байесовского подхода в тематическом моделировании и разобщённость академического и индустриального сообщества приводит к распространению некоторых заблуждений, достойных упоминания и критики.

- «Тематическое моделирование — это в основном LDA». Нет, есть сотни моделей, решающих разнообразные задачи, с которыми плохо справляется LDA.
- «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 предсказывают частоты слов в документах». Формально да, но их цель в другом — выявление 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, регуляризаторы E-шага позволяют учитывать порядок слов в документах.
- «Тематические векторы не отражают смысл слов, как word2vec». Это так в обычных моделях, которые строятся по частотам слов в документах. Но можно строить тематические модели по частотам парных сочетаний слов, как и word2vec. Тематические векторы в моделях битермов (BitermTM) или сети слов (WNTM) не только определяют семантическую близость слов, но и обладают свойствами интерпретируемости и разреженности.
- «Тематическая модель LDA переобучается гораздо меньше, чем PLSA». Нет, их качество примерно одинаково, особенно на больших коллекциях. LDA осторожнее оценивает вероятности редких слов. Формально это улучшает правдоподобие, хотя 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.

Некоторые проблемы пока остаются открытыми в ARTM. Как подбирать коэффициенты регуляризации адаптивно и полностью автоматически в ходе итераций? Как решить проблему несбалансированности тем, когда крупные темы разделяются и образуют темы-дубликаты, а мелкие объединяются и образуют мусорные темы? Как обеспечить построение полного набора хорошо интерпретируемых тем?

В данный обзор не вошли некоторые важные типы моделей, например, для анализа изображений и видеопотоков, аннотирования изображений, рекомендательных систем, суммаризации текстов, анализа тональности и выявления мнений, использования и построения онтологий, обнаружения новых тем и прослеживания новостных сюжетов. Также не были затронуты вопросы инициализации и автоматического именования тем. Не нашла отражения в обзоре новая тенденция создания гибридных моделей на основе тематического моделирования и нейронных сетей.

Обзор написан по материалам спецкурса «Вероятностное тематическое моделирование»[11], который автор читает на факультете ВМК Московского Государственного Университета им. М. В. Ломоносова и на кафедре интеллектуальных систем Московского Физико-Технического Института. Обновляемая электронная версия доступна на сайте MachineLearning.ru[12].

Работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований (проекты 17-07-01536, 20-07-00936) и правительства Российской Федерации (соглашение 05.Y09.21.0018).

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] *Воронцов К. В.* Аддитивная регуляризация тематических моделей коллекций текстовых документов // *Доклады РАН.* — 2014. — Т. 456, № 3. — С. 268–271.

[5] *Воронцов К. В., Потапенко А. А.* Регуляризация, робастность и разреженность вероятностных тематических моделей // *Компьютерные исследования и моделирование.* — 2012. — Т. 4, № 4. — С. 693–706.

[6] *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.

[7] *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.

[8] *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.

[9] *Дударенко М. А.* Регуляризация многоязычных тематических моделей // *Вычислительные методы и программирование.* — 2015. — Т. 16. — С. 26–38.

[10] *Ирхин И. А., Воронцов К. В.* Аддитивная регуляризация тематических моделей с быстрой векторизацией текста // *ЖВМиМФ.* — 2020. — Т. ??, № ?? — С. ??–??

[11] *Ирхин И. А., Воронцов К. В.* Сходимость алгоритма аддитивной регуляризации тематических моделей // *Труды Института математики и механики УрО РАН.* — 2020. — Т. ??, № ?? — С. ??–??

[12] *Лукашевич Н. В.* Тезаурусы в задачах информационного поиска. — Издательство МГУ имени М. В. Ломоносова, 2011.

[11] `http://www.MachineLearning.ru/wiki?title=BTM`.

[12] `http://www.MachineLearning.ru/wiki/images/d/d5/Voron17survey-artm.pdf`.

[13] *Маннинг К. Д., Рагхаван П., Шютце Х.* Введение в информационный поиск. — Вильямс, 2011.

[14] *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.

[15] *Тихонов А. Н., Арсенин В. Я.* Методы решения некорректных задач. — М.: Наука, 1986.

[16] *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.

[17] *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.

[18] *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.

[19] *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.

[20] *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.

[21] *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.

[22] *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.

[23] *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.

[24] *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.

[25] *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.

[26] *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.

[27] *Bishop C. M.* Pattern Recognition and Machine Learning. — Springer, Series: Information Science and Statistics, 2006. — 740 pp.

[28] *Blei D., Lafferty J.* A correlated topic model of Science // *Annals of Applied Statistics.* — 2007. — Vol. 1. — Pp. 17–35.

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

[30] *Blei D. M., Griffiths T., Jordan M., Tenenbaum J.* Hierarchical topic models and the nested chinese restaurant process // NIPS. — 2003.

[31] *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.

[32] *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.

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

[34] *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.

[35] *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.

[36] *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.

[37] *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.

[38] *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.

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

[40] *Chien J.-T., Chang Y.-L.* Bayesian sparse topic model // *Journal of Signal Processessing Systems.* — 2013. — Vol. 74. — Pp. 375–389.

[41] *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.

[42] *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.

[43] *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.

[44] *Dai A. M., Olah C., Le Q. V.* Document embedding with paragraph vectors // NIPS Deep Learning Workshop. — 2015.

[45] *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.

[46] *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.

[47] *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.

[48] *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.

[49] *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.

[50] *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.

[51] *Eisenstein J., Ahmed A., Xing E. P.* Sparse additive generative models of text // ICML'11. — 2011. — Pp. 1041–1048.

[52] *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.

[53] *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.

[54] *Feldman S. E.* The answer machine // Synthesis Lectures on Information Concepts, Retrieval, and Services. — Morgan & Claypool Publishers, 2012. — Vol. 4. — Pp. 1–137.

[55] *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.

[56] *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.

[57] *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.

[58] *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.

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

[60] *Hoffman M. D., Blei D. M., Bach F. R.* Online learning for latent Dirichlet allocation // NIPS. — Curran Associates, Inc., 2010. — Pp. 856–864.

[61] *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.

[62] *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.

[63] *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.

[64] *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.

[65] *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.

[66] *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.

[67] *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.

[68] *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.

[69] *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.

[70] *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.

[71] *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.

[72] *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.

[73] *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.

[74] *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.

[75] *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.

[76] *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.

[77] *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.

[78] *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.

[79] *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.

[80] *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.

[81] *Li W., McCallum A.* Pachinko allocation: Dag-structured mixture models of topic correlations // ICML. — 2006.

[82] *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.

[83] *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.

[84] *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.

[85] *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.

[86] *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.

[87] *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.

[88] *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.

[89] *Marchionini G.* Exploratory search: From finding to understanding // *Commun. ACM*. — 2006. — Vol. 49, no. 4. — Pp. 41–46.

[90] *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.

[91] *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.

[92] *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.

[93] *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.

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

[95] *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.

[96] *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.

[97] *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.

[98] *Mimno D., Li W., McCallum A.* Mixtures of hierarchical topics with pachinko allocation // ICML. — 2007.

[99] *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.

[100] *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.

[101] *Minka T. P.* Estimating a Dirichlet distribution: Tech. rep.: 2000 (revised 2003, 2009, 2012).

[102] *Moretti F.* Graphs, maps, trees : abstract models for literary history. — London; New York: Verso, 2007.

[103] *Nadeau D., Sekine S.* A survey of named entity recognition and classification // *Linguisticae Investigationes*. — 2007. — Vol. 30, no. 1. — Pp. 3–26.

[104] *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.

[105] *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.

[106] *Newman D., Karimi S., Cavedon L.* External evaluation of topic models // Australasian Document Computing Symposium. — December 2009. — Pp. 11–18.

[107] *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.

[108] *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.

[109] *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.

[110] *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.

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

[112] *Pagliardini M., Gupta P., Jaggi M.* Unsupervised learning of sentence embeddings using compositional n-gram features // *CoRR*. — 2017. — Vol. abs/1703.02507.

[113] *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.

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

[115] *Paul M. J., Girju R.* Topic modeling of research fields: An interdisciplinary perspective // RANLP. — RANLP 2009 Organising Committee / ACL, 2009. — Pp. 337–342.

[116] *Pennington J., Socher R., Manning C. D.* GloVe: Global vectors for word representation // Empirical Methods in Natural Language Processing (EMNLP). — 2014. — Pp. 1532–1543.

[117] *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.

[118] *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.

[119] *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.

[120] *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.

[121] *Pritchard J. K., Stephens M., Donnelly P.* Inference of population structure using multilocus genotype data // *Genetics*. — 2000. — Vol. 155. — Pp. 945–959.

[122] *Pujara J., Skomoroch P.* Large-scale hierarchical topic models // NIPS Workshop on Big Learning. — 2012.

[123] *Rahman M.* Search engines going beyond keyword search: A survey // *International Journal of Computer Applications*. — August 2013. — Vol. 75, no. 17. — Pp. 1–8.

[124] *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.

[125] *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.

[126] *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.

[127] *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.

[128] *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.

[129] *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.

[130] *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.

[131] *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.

[132] *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.

[133] *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.

[134] *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.

[135] *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.

[136] *Si X., Sun M.* Tag-LDA for scalable real-time tag recommendation // *Journal of Information & Computational Science*. — 2009. — Vol. 6. — Pp. 23–31.

[137] *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.

[138] *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.

[139] *Steyvers M., Griffiths T.* Finding scientific topics // *Proceedings of the National Academy of Sciences.* — 2004. — Vol. 101, no. Suppl. 1. — Pp. 5228–5235.

[140] *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.

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

[142] *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.

[143] *Teh Y. W., Newman D., Welling M.* A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation // NIPS. — 2006. — Pp. 1353–1360.

[144] 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.

[145] *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.

[146] *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.

[147] *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.

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

[149] *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.

[150] *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.

[151] *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.

[152] *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.

[153] *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.

[154] *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.

[155] *Wallach H.* Structured Topic Models for Language: Ph.D. thesis / Newnham College, University of Cambridge. — 2008.

[156] *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.

[157] *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.

[158] *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.

[159] *Wang C., Blei D. M.* Decoupling sparsity and smoothness in the discrete hierarchical Dirichlet process // NIPS. — Curran Associates, Inc., 2009. — Pp. 1982–1989.

[160] *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.

[161] *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.

[162] *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.

[163] *Wang C., Liu X., Song Y., Han J.* Scalable and robust construction of topical hierarchies // *CoRR.* — 2014. — Vol. abs/1403.3460.

[164] *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.

[165] *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.

[166] *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.

[167] *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.

[168] *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.

[169] *Wu L. Y., Fisch A., Chopra S., Adams K., Bordes A., Weston J.* Starspace: Embed all the things! // Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA, February 2–7, 2018. — 2018. — Pp. 5569–5577.

[170] *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.

[171] *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.

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

[173] *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.

[174] *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.

[175] *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.

[176] *Yin H., Cui B., Sun Y., Hu Z., Chen L.* LCARS: A spatial item recommender system // *ACM Transaction on Information Systems*. — 2014.

[177] *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.

[178] *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.

[179] *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.

[180] *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.

[181] *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.

[182] *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.

[183] *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.