

Вероятностное тематическое моделирование: обзор моделей и аддитивная регуляризация

Воронцов Константин Вячеславович

vokov@forecsys.ru

Московский физико-технический институт (государственный университет),
Федеральный исследовательский центр «Информатика и управление» РАН

19 мая 2018 г.

Содержание

1	Введение	4
2	Основы тематического моделирования	6
	Предварительная обработка текста.	6
	Гипотеза о существовании тем.	6
	Гипотеза «мешка слов».	7
	Гипотеза о вероятностном порождении данных.	7
	Гипотеза условной независимости.	7
	Вероятностная тематическая модель порождения текста.	7
	Задача тематического моделирования	8
	Низкоранговое матричное разложение.	8
	Частотные оценки условных вероятностей.	9
	EM-алгоритм.	9
	Рациональный EM-алгоритм.	10
3	Аддитивная регуляризация	10
	Принцип максимума правдоподобия	10
	Регуляризация.	11
	Необходимые условия максимума.	11
	Модель вероятностного латентного семантического анализа	13
	Регуляризованный EM-алгоритм	13
	Онлайновый EM-алгоритм	14
	О стратегиях регуляризации.	14
	Относительные коэффициенты регуляризации.	14
4	Байесовская регуляризация и модель LDA	16
	Байесовская регуляризация.	16
	Модель латентного размещения Дирихле.	17
	Дивергенция Кульбака–Лейблера	18

Не-вероятностная интерпретация модели LDA.	19
5 Интерпретируемость тем	20
Сглаживание и разреживание.	20
Частичное обучение.	20
Предметные и фоновые темы.	21
Сфокусированный тематический поиск.	21
Декоррелирование.	22
Комбинирование регуляризаторов	22
6 Моделирование мультимодальных данных	23
Мультязычные модели.	24
Модальности категорий и авторов.	26
Темпоральные модели.	27
7 Моделирование транзакционных данных	29
Тематические модели на гиперграфах.	29
8 Моделирование зависимостей	33
Классификация.	33
Регрессия.	34
Корреляции тем.	35
9 Определение числа тем	35
10 Построение тематических иерархий	36
Регуляризатор межуровневых связей.	37
Регуляризатор разреживания межуровневых связей	38
11 Моделирование связей между документами	38
Ссылки и цитирование.	38
Геолокации.	39
Графы и социальные сети.	40
12 Моделирование встречаемости слов	41
Коллокации.	41
Битермы.	42
Сеть слов.	43
Когерентность.	44
Модели векторных представлений слов	45
13 Тематическая сегментация текстов	46
Тематическая модель предложений.	46
Тематическая модель сегментации.	47
Регуляризатор E-шага.	48
Разреживание распределений $p(t d, w)$	50
Разреживающий регуляризатор E-шага для сегментации.	51

14 Критерии качества тематических моделей	51
Внешние критерии	51
Перспексия.	52
Когерентность.	52
Разреженность и различность тем.	53
15 Проект VigARTM	54
Подготовка данных.	55
Словари VigARTM	55
Регуляризаторы	56
Метрики качества	58
Выгрузка параметров модели.	59
Многопоточный пакетный EM-алгоритм.	59
Иерархическая тематическая модель	61
16 Разведочный информационный поиск	61
Тематический поиск.	62
Качество разведочного поиска.	63
Визуализация.	63
17 Заключение	64

1 Введение

Тематическое моделирование — одно из современных направлений статистической *обработки естественного языка* (natural language processing, NLP), активно развивающееся с конца 90-х годов. Тематическая модель выявляет тематику коллекции текстовых документов, определяя, к каким темам относится каждый документ, и какие слова или фразы образуют каждую тему. В хорошей модели каждая тема является семантически однородной, то есть несёт в себе определённый «атомарный смысл». 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. Текст произвольной длины может быть преобразован в тематическое векторное представление, из которого можно узнать, какая доля каждой из тем (атомарных смыслов) содержится в данном тексте. Похожую задачу решают модели *векторных представлений слов* (word embedding) [79, 65], предложений [95, 149] и документов [35], однако в них координаты векторов не имеют смысловой интерпретации.

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. Многочисленные слова и омонимы, наоборот, распределяют свои вероятности по нескольким семантически не связанным темам. Например, значение слова «ядро» может быть понято из того, какая тема доминирует в его контексте — математика, физика, биология или военная история.

Многие приложения текстовой аналитики используют тематические векторные представления текста: выявление трендов в новостных потоках, патентных базах, архивах научных публикаций [159, 126], многоязычный информационный поиск [137, 136], поиск тематических сообществ в социальных сетях [161, 128, 101, 27], классификация и категоризация документов [111, 162], тематическая сегментация текстов [145], тегирование веб-страниц [60], обнаружение текстового спама [10]. Существуют и не-текстовые приложения тематического моделирования в анализе изображений и видеопотоков [50, 68, 43, 127], в рекомендательных системах [153, 140, 64, 156, 155], в популяционной генетике [104], в биоинформатике для анализа нуклеотидных [61] и аминокислотных последовательностей [116, 59]. Другие приложения тематических моделей упоминаются в обзорах [36, 22].

Построение тематической модели по коллекции документов является некорректно поставленной оптимизационной задачей, которая может иметь бесконечное множество решений. Согласно теории регуляризации А. Н. Тихонова [11], решение такой задачи возможно доопределить и сделать устойчивым. Для этого к основному критерию добавляется *регуляризатор* — дополнительный критерий, учитывающий специ-

фические 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 [3, 132]. 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 обобщённый EM-алгоритм, к которому можно подключать регуляризаторы как модули [130, 44, 57]. Идея модульности реализована в библиотеке тематического моделирования с открытым кодом **BigARTM** (<http://bigartm.org>). 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) [26].

Основная цель данного обзора — показать разнообразие тематических моделей, сосредоточившись на первом и 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, даже не скрывая математических выкладок. Сопоставимый по охвату и обстоятельности обзор байесовских моделей занял бы сотни страниц.

Разделы 2–4 являются базовыми. В разделах 5–13 в терминах регуляризации описываются различные виды тематических моделей. Эти разделы практически не связаны друг с другом, их можно читать в произвольном порядке или использовать как путеводитель по ссылкам на литературу. В разделе 7 предлагается новое и наиболее широкое обобщение тематических моделей для обработки транзакционных данных произвольной природы. Раздел 12 посвящён тематическим моделям дистрибутивной семантики, успешно конкурирующим с векторными представлениями слов семейства word2vec. В разделе 13 предлагается новый тип регуляризации, позволяющий учитывать порядок слов в документах в обход гипотезы «мешка слов». Раздел 14 посвящён оцениванию качества тематических моделей. Раздел 15 содержит начальные сведения об использовании библиотеки **BigARTM**. В разделе 16 обсуждается применение тематического моделирования для разведочного информационного поиска. В разделе 17 — краткое заключение.

2 Основы тематического моделирования

В этом разделе будут введены основные понятия и постановка задачи тематического моделирования. Будет показан простой эвристический способ её решения.

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 `rumorphy`.

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. Они бесполезны для тематического моделирования и могут быть отброшены. К ним относятся предлоги, союзы, числительные, местоимения, некоторые глаголы, прилагательные и наречия. Число таких слов обычно варьируется в пределах нескольких сотен. Их отбрасывание почти не влияет на объём словаря, но может приводить к заметному сокращению длины некоторых 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 — словосочетаний, характерных для предметной области. Их использование вместо отдельных слов или наряду с ними улучшает интерпретируемость тем. Для их выделения можно использовать тезаурусы [8] или методы автоматического выделения терминов (automatic term extraction, АТЕ), не требующие привлечения экспертов [41, 69, 113].

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 [86, 62, 92].

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, \dots, w_{n_d} из словаря W .

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

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

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

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

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

равно вероятности того, что 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. В большинстве случаев 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 может полагаться 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) , $i = 1, \dots, n$;

- 1 $i := 0$;
 - 2 для всех $d \in D$
 - 3 для всех $j = 1, \dots, n_d$
 - 4 $i := i + 1$; $d_i := d$;
 - 5 выбрать случайную тему t_i из распределения $p(t|d_i)$;
 - 6 выбрать случайный терм w_i из распределения $p(w|t_i)$;
-

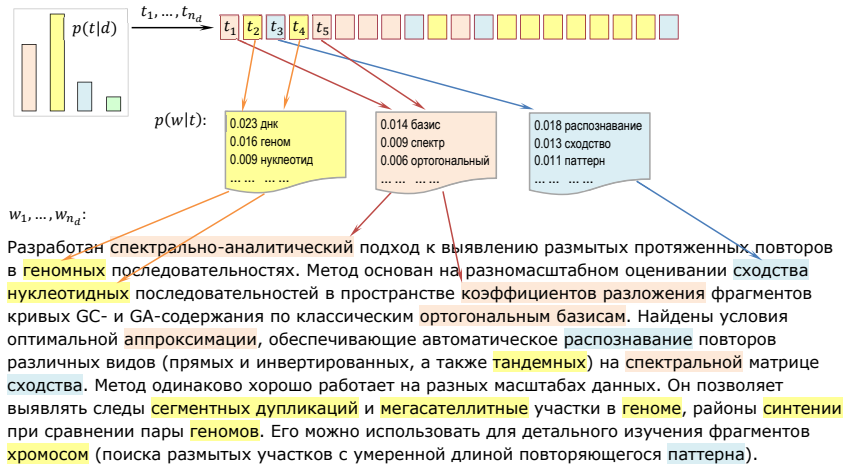


Рис. 1: Процесс порождения текстовой коллекции вероятностной тематической моделью (2): в каждой позиции i документа d_i сначала порождается тема $t_i \sim p(t|d_i)$, затем терм $w_i \sim p(w|t_i)$.

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 определение тематики документов и связанных с ними объектов. Также требуется находить распределения $\varphi_{wt} = p(w|t)$, описывающие семантику каждой темы t словами естественного языка.

Низкоранговое матричное разложение. Равенство (2) можно переписать в матричном виде. В левой части равенства находится известная матрица частот термов в документах $F = (\hat{p}(w|d))_{W \times D}$. 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 $\Phi\Theta$ будем называть *стохастическим матричным разложением*.

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

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

n_{dw} — число вхождений термина w в документ d ;

$n_d = \sum_w n_{dw}$ — длина документа d в терминах;

$n_w = \sum_d n_{dw}$ — число вхождений термина w во все документы коллекции;

$n = \sum_d \sum_w n_{dw}$ — длина коллекции в терминах.

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

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

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 $n \rightarrow \infty$ 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:

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

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

$$p_{tdw} = \frac{\varphi_{wt}\theta_{td}}{\sum_s \varphi_{ws}\theta_{sd}}; \quad (5)$$

$$\varphi_{wt} = \frac{n_{wt}}{\sum_{w'} n_{w't}}; \quad n_{wt} = \sum_{d \in D} n_{dw} p_{tdw}; \quad (6)$$

$$\theta_{td} = \frac{n_{td}}{\sum_{t'} n_{t'd}}; \quad n_{td} = \sum_{w \in d} n_{dw} p_{tdw}. \quad (7)$$

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 (5)–(7) продолжаются в цикле до сходимости.

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

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

Выход: параметры модели Φ и Θ ;

1 **повторять**

2 обнулить n_{wt} , n_{td} , n_t для всех $d \in D$, $w \in W$, $t \in T$;

3 **для всех** $d \in D$, $w \in d$

4 $n_{tdw} := n_{dw} \varphi_{wt} \theta_{td} / \sum_{\tau} \varphi_{w\tau} \theta_{\tau d}$ для всех $t \in T$;

5 увеличить n_{wt} , n_{td} , n_t на n_{tdw} для всех $t \in T$;

6 $\varphi_{wt} := n_{wt} / n_t$ для всех $w \in W$, $t \in T$;

7 $\theta_{td} := n_{td} / n_d$ для всех $d \in D$, $t \in T$;

8 **пока** Φ и Θ не сойдутся;

Этот итерационный процесс является частным случаем EM-алгоритма, предназначенного для построения вероятностных моделей со скрытыми переменными [38]. Вычисление условных распределений скрытых переменных (5) называется E-шагом (expectation), вычисление параметров модели (6)–(7) — M-шагом (maximization).

Далее мы выведем EM-алгоритм из общей оптимизационной постановки задачи. Сейчас мы пришли к нему элементарным путём, который даёт простое интуитивное понимание сути EM-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 EM-алгоритм. Вычисление переменных n_{wt} , n_{td} , n_t на M-шаге требует однократного прохода коллекции в цикле по всем термам $w \in d$ всех документов $d \in D$. Внутри этого цикла каждая переменная p_{tdw} вычисляется только один раз. Поэтому E-шаг встраивается внутрь M-шага без дополнительных вычислительных затрат и без хранения трёхмерной матрицы p_{tdw} . Этот вариант реализации EM-алгоритма будем называть *рациональным*; он показан в алгоритме 2.

3 Аддитивная регуляризация

В этом разделе вводится общий формализм оптимизации и регуляризации, который далее будет использоваться для описания всех тематических моделей.

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 выборки от параметров модели. Благодаря предположению о независимости наблюдений, она равна произведению вероятностей термов в документах:

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

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

$$\sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \varphi_{wt} \theta_{td} \rightarrow \max_{\Phi, \Theta}; \quad (8)$$

$$\sum_{w \in W} \varphi_{wt} = 1; \quad \varphi_{wt} \geq 0; \quad \sum_{t \in T} \theta_{td} = 1; \quad \theta_{td} \geq 0. \quad (9)$$

Регуляризация. 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* [11]. 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) [3] основана на максимизации линейной комбинации логарифма правдоподобия и *регуляризаторов* $R_i(\Phi, \Theta)$ с неотрицательными *коэффициентами регуляризации* τ_i , $i = 1, \dots, k$:

$$\sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \varphi_{wt} \theta_{td} + R(\Phi, \Theta) \rightarrow \max_{\Phi, \Theta}; \quad R(\Phi, \Theta) = \sum_{i=1}^k \tau_i R_i(\Phi, \Theta); \quad (10)$$

при ограничениях неотрицательности и нормировки

$$\sum_{w \in W} \varphi_{wt} \in \{0, 1\}; \quad \varphi_{wt} \geq 0; \quad \sum_{t \in T} \theta_{td} \in \{0, 1\}; \quad \theta_{td} \geq 0. \quad (11)$$

Условия нормировки с альтернативной правой частью (11) ослабляют условия (9), допуская появление нулевых столбцов в матрицах решения Φ и Θ .

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

Необходимые условия максимума. Задача (10), (9) относится к классу невыпуклых задач математического программирования. Для неё возможно найти лишь локальный экстремум, качество которого зависит от начального приближения. На практике поиск более глубокого экстремума не столь важен, как адекватная формализация дополнительных критериев и поиск компромисса между ними.

Введём оператор norm , который преобразует произвольный заданный вектор $(x_i)_{i \in I}$ в вектор вероятностей $(p_i)_{i \in I}$ дискретного распределения путём обнуления отрицательных элементов и нормировки:

$$p_i = \text{norm}_{i \in I}(x_i) = \frac{(x_i)_+}{\sum_{j \in I} (x_j)_+}, \quad \text{для всех } i \in I,$$

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

Теорема 1. Пусть функция $R(\Phi, \Theta)$ непрерывно дифференцируема. Тогда точка (Φ, Θ) локального экстремума задачи (10), (11) удовлетворяет системе уравнений со вспомогательными переменными $p_{tdw} = p(t|d, w)$:

$$p_{tdw} = \text{norm}_{t \in T}(\varphi_{wt}\theta_{td}); \quad (12)$$

$$\varphi_{wt} = \text{norm}_{w \in W}\left(n_{wt} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}}\right); \quad n_{wt} = \sum_{d \in D} n_{dw}p_{tdw}; \quad (13)$$

$$\theta_{td} = \text{norm}_{t \in T}\left(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right); \quad n_{td} = \sum_{w \in W} n_{dw}p_{tdw}. \quad (14)$$

Доказательство. Запишем необходимые условия Каруша–Куна–Таккера локального экстремума задачи (10), (9):

$$\sum_{d \in D} n_{dw} \frac{\theta_{td}}{p(w|d)} + \frac{\partial R}{\partial \varphi_{wt}} = \lambda_t - \lambda_{wt}; \quad \lambda_{wt} \geq 0; \quad \lambda_{wt}\varphi_{wt} = 0; \quad (15)$$

$$\sum_{w \in W} n_{dw} \frac{\varphi_{wt}}{p(w|d)} + \frac{\partial R}{\partial \theta_{td}} = \mu_d - \mu_{td}; \quad \mu_{td} \geq 0; \quad \mu_{td}\theta_{td} = 0; \quad (16)$$

где множители Лагранжа λ_t, μ_d соответствуют ограничениям нормировки, множители λ_{wt}, μ_{td} — ограничениям неотрицательности.

Умножим обе части равенства (15) на φ_{wt} , обе части равенства (16) на θ_{td} , и выделим вспомогательные переменные p_{tdw}, n_{wt} и n_{td} :

$$\varphi_{wt}\lambda_t = \sum_{d \in D} n_{dw} \frac{\varphi_{wt}\theta_{td}}{p(w|d)} + \varphi_{wt} \frac{\partial R}{\partial \varphi_{wt}} = n_{wt} + \varphi_{wt} \frac{\partial R}{\partial \varphi_{wt}};$$

$$\theta_{td}\mu_d = \sum_{w \in W} n_{dw} \frac{\varphi_{wt}\theta_{td}}{p(w|d)} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} = n_{td} + \theta_{td} \frac{\partial R}{\partial \theta_{td}}.$$

Если $\lambda_t \leq 0$, то $\varphi_{wt} = 0$ для всех $w \in W$. В этом случае тема t исключается из модели, что допускается ослабленным условием нормировки (11).

Если $\lambda_t > 0$, то либо обе части равенства положительны, либо $\varphi_{wt} = 0$. Объединим эти два случая в одну формулу:

$$\varphi_{wt}\lambda_t = \left(n_{wt} + \varphi_{wt} \frac{\partial R}{\partial \varphi_{wt}}\right)_+. \quad (17)$$

Если $\mu_d \leq 0$, то $\theta_{td} = 0$ для всех $t \in T$. В этом случае документ d исключается из модели, что допускается ослабленным условием нормировки (11).

Если $\mu_d > 0$, то либо $\theta_{td} = 0$, либо обе части равенства положительны. Объединим эти два случая в одну формулу:

$$\theta_{td}\mu_d = \left(n_{td} + \theta_{td} \frac{\partial R}{\partial \theta_{td}}\right)_+. \quad (18)$$

Суммируем левую и правую части равенства (17) по w , равенства (18) по t :

$$\lambda_t = \sum_{w \in W} \left(n_{wt} + \varphi_{wt} \frac{\partial R}{\partial \varphi_{wt}} \right)_+; \quad (19)$$

$$\mu_d = \sum_{t \in T} \left(n_{td} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right)_+. \quad (20)$$

Подставляя λ_t из (19) в (17), получим (13).

Подставляя μ_d из (20) в (18), получим (14).

Теорема доказана.

Будем называть модель *вырожденной по теме t* , если $\varphi_{wt} = 0$ для всех $w \in W$, и *вырожденной по документу d* , если $\theta_{td} = 0$ для всех $t \in T$. Вырожденность возникает в случаях, когда регуляризатор R оказывает чрезмерное разреживающее воздействие на параметры модели. Вырожденные темы и документы приходится исключать из модели. На практике вырожденность возникает довольно редко. Сокращение числа тем может быть желательным побочным эффектом регуляризации. Вырожденность документа может означать, что модель не в состоянии его описать, например, если он слишком короткий или не соответствует тематике коллекции.

Модель вероятностного латентного семантического анализа (probabilistic latent semantic analysis, PLSA) — это первая вероятностная тематическая модель, предложенная Томасом Хофманном в 1999 году [49]. В ARTM она соответствует нулевому регуляризатору, $R(\Phi, \Theta) = 0$. В этом случае система (12)–(14) совпадает с системой (5)–(7), которую мы получили ранее из элементарных соображений. Вычисление E-шага по формуле Байеса (5) остаётся при любой регуляризации. В модели PLSA не может быть вырожденных тем или документов.

Регуляризованный EM-алгоритм является применением метода простых итераций для решения системы (12)–(14). Сначала выбираются начальные приближения $\varphi_{wt}, \theta_{td}$, затем в цикле до сходимости чередуются *E-шаг* (12) и *M-шаг* (13)–(14). Рациональный вариант регуляризованного EM-алгоритма строится аналогично алгоритму 2, только шаги 6 и 7 заменяются формулами M-шага (13)–(14).

Известно, что EM-алгоритм без регуляризации сходится в слабом смысле: на каждой итерации правдоподобие увеличивается [38]. Аналогичные условия слабой сходимости для ARTM получены И. А. Ирхиным¹.

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 [5].

В *стохастическом EM-алгоритме* [5] для каждого термина w_i семплируется только одна тема $t_i \sim p(t|d, w_i)$, затем на M-шаге вместо $p(t|d, w_i)$ подставляется вырожденное распределение $\hat{p}(t|d, w_i) = [t = t_i]$. Эта идея эксплуатируется также в байесовском обучении тематических моделей на основе семплирования Гиббса [121].

EM-подобные алгоритмы тематического моделирования сравниваются в [18].

¹Ирхин И. А. Сходимость численных методов вероятностного тематического моделирования. Магистерская диссертация, МФТИ, 2016.

<http://www.MachineLearning.ru/wiki/images/0/05/Irkhin2016msc.pdf>

Онлайновый EM-алгоритм считается наиболее быстрым и хорошо распараллеливается [48, 20]. 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 онлайн-алгоритмы способны обрабатывать потоковые данные.

Детали параллельной реализации оффлайн- и онлайн-EM-алгоритма в библиотеке `BigARTM` описаны в разделе 15, ещё подробнее — в статье [44].

О стратегиях регуляризации. Задача тематического моделирования по сути является многокритериальной. Темы должны удовлетворять многим требованиям одновременно: интерпретируемости, различности, разреженности и т. д. Кроме того, тематическая модель обычно используется не сама по себе, а как вспомогательный инструмент для решения различных 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, поскольку ничто не запрещает изменять коэффициенты τ_i в ходе итераций. Одни регуляризаторы могут делать подготовительную работу для других. Некоторые регуляризаторы рекомендуется включать, когда EM-алгоритм уже начал сходиться. Другие лучше отключать после того, как они выполнили свою работу. Некоторые регуляризаторы могут нейтрализовать друг друга, и тогда их приходится применять поочередно. Регуляризаторы, как яды для модели, — в малых дозах они лечат, в больших они смертельно опасны, в сочетании друг с другом они могут давать неожиданные эффекты. Систематизация и использование этих эффектов становится предметом исследования в ARTM.

Стратегией регуляризации называется функция коэффициентов регуляризации τ_i от номера итерации. В частности, она может использовать текущие значения параметров модели и метрик качества.

Относительные коэффициенты регуляризации. Ещё одна проблема заключается в том, что коэффициенты регуляризации, тщательно подобранные для одной коллекции, скорее всего, будут плохо работать на другой. Они могут зависеть от размера коллекции, мощности словаря, средней длины документов. Если коллекция пополняется, то со временем может потребоваться их перенастройка. Проблема решается введением нормировочных коэффициентов и *относительных коэффициентов регуляризации*, выражающих степень воздействия регуляризатора на тематическую модель. Относительные коэффициенты меньше зависят от коллекции, и для них могут быть установлены универсальные рекомендуемые значения.

Рассмотрим формулу М-шага (13) со взвешенной суммой регуляризаторов R_i :

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

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

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

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 $\tilde{\tau}_i \rightarrow 0$ регуляризатор R_i отключается. При $\tilde{\tau}_i \rightarrow \infty$ 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 (14) 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:

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

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 Байесовская регуляризация и модель LDA

Байесовская регуляризация. До сих пор мы предполагали, что данные порождаются вероятностной моделью с параметрами (Φ, Θ) , которые не известны и не случайны. В байесовском подходе предполагается, что параметры также случайны и подчиняются некоторому *априорному* распределению $p(\Phi, \Theta; \gamma)$ с неслучайным *гиперпараметром* γ . В этом случае максимизация совместного правдоподобия данных и модели приводит к принципу *максимума апостериорной вероятности* (maximum a posteriori probability, MAP):

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

После логарифмирования получаем модификацию задачи (8), в которой логарифм априорного распределения является регуляризатором:

$$\sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \varphi_{wt} \theta_{td} + \underbrace{\ln p(\Phi, \Theta; \gamma)}_{R(\Phi, \Theta)} \rightarrow \max_{\Phi, \Theta, \gamma}. \quad (21)$$

В байесовском подходе применяется также принцип максимизации неполного правдоподобия, в котором по случайным параметрам (Φ, Θ) производится интегрирование и оптимизируются гиперпараметры γ . Считается, что этот приём снижает размерность задачи и риск переобучения. Действительно, размерность вектора γ , как правило, много меньше размеров матриц Φ, Θ и не зависит от объёма коллекции. Однако для решения прикладных задач всё равно нужны именно эти матрицы. Формулы для них выводятся громоздкими приближёнными методами, но в итоге мало отличаются от MAP-оценок [18].

В байесовском подходе оцениваются не сами параметры Φ, Θ , а их апостериорное распределение $p(\Phi, \Theta | D; \gamma)$. Для задач тематического моделирования в этом нет особой необходимости. На практике полученное распределение используется исключительно для того, чтобы вернуться к точечным оценкам математического ожидания. Ни апостериорные распределения, ни интервальные оценки, ни даже альтернативные точечные оценки медианы или моды в приложениях не используются.

Техники приближённого байесовского вывода (вариационный вывод [125], сэмплирование Гиббса [121], распространение ожидания) не позволяют легко комбинировать модели и добавлять регуляризаторы, не имеющие вероятностной интерпретации. Для каждой новой модели приходится заново выполнять математические выкладки и программную реализацию. В прикладных проектах сроки, стоимость и риски таких разработок становятся непреодолимым барьером. Поэтому на практике пользуются простой устаревшей моделью LDA, а байесовское тематическое моделирование редко выходит за рамки академических исследований. Тем не менее, в литературе по тематическому моделированию байесовский подход доминирует.

Многокритериальный не-байесовский подход ARTM — это попытка изменить ситуацию. Байесовские тематические модели в большинстве случаев удаётся переформулировать в терминах регуляризации, записав постановку задачи в виде (21). С этого момента регуляризатор отделяется от модели и может быть использован в других моделях. Это приводит к модульной технологии тематического моделирования, которая реализована и развивается в проекте BigARTM.

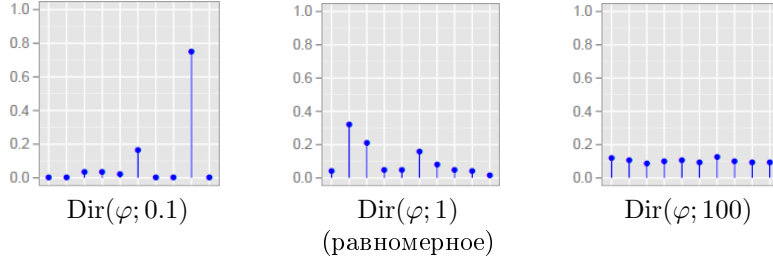


Рис. 2: Пример трёх неотрицательных нормированных векторов $\varphi_t \in \mathbb{R}^{10}$, порождённых соответственно тремя симметричными распределениями Дирихле с параметрами 0.1, 1, 100.

Модель латентного размещения Дирихле. Дэвид Блэй, Эндрю Ён и Майкл Джордан предложили модель LDA (latent Dirichlet allocation) [26] для решения проблемы переобучения в PLSA, которая предсказывала вероятности термов $p(w|d)$ на новых документах заметно хуже, чем на обучающей коллекции. Позже выяснилось, что на больших коллекциях обе модели почти не переобучаются, а их правдоподобия отличаются незначительно [76, 150, 71]. Различия проявляются только на низкочастотных термах, которые не важны для образования тем. В робастных вариантах PLSA и LDA такие термы игнорируются, что резко снижает переобучение, так и различие в правдоподобии моделей [103]. Сам вопрос о переобучении поставлен не вполне корректно. Во-первых, тематические модели строятся не ради предсказания термов в документах, а для выявления кластерной структуры коллекции. Во-вторых, величина переобучения зависит не только от самой модели, но и от того, как мы договоримся измерять её качество. Для измерения обычно используется перплексия, которая сильно штрафует заниженные вероятности низкочастотных термов. Возможно, превосходство LDA над PLSA не столь существенно, как это принято считать. Тем не менее, LDA де факто является самой используемой тематической моделью.

Модель LDA основана на предположении, что столбцы θ_d и φ_t являются случайными векторами, которые порождаются распределениями Дирихле с параметрами $\alpha \in \mathbb{R}^{|T|}$ и $\beta \in \mathbb{R}^{|W|}$ соответственно:

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

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

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

Распределения Дирихле способны породить как разреженные, так и плотные векторы дискретных распределений, рис.2. Чем меньше β_w , тем более разрежена соответствующая w компонента φ_{wt} в порождаемых векторах φ_t . 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.

Тематическая модель порождения данных является двухуровневой: сначала из распределения Дирихле порождаются вектор-столбцы φ_t , которые задают темы.

Затем из полученных распределений $p(w|t) = \varphi_{wt}$ порождаются термы, образующие тематические части документов $p(w|t, d)$. Таким образом, двухуровневая модель описывает кластерные структуры в текстовых коллекциях. Векторы распределений $p(w|t)$ интерпретируются как центры кластеров, а распределения $p(w|t, d)$ являются точками этих кластеров.

Более убедительных лингвистических обоснований распределение Дирихле не имеет. Его широкое распространение в тематическом моделировании объясняется скорее его чисто математическим удобством для байесовского обучения. Распределение Дирихле является сопряжённым к мультиномиальному распределению, что существенно упрощает байесовский вывод. Благодаря этому свойству оно оказывается «на особом положении» в байесовском тематическом моделировании, и большинство моделей строятся с использованием распределений Дирихле.

Согласно (21), модели LDA соответствует регуляризатор, с точностью до константы равный логарифму априорного распределения Дирихле:

$$\begin{aligned} R(\Phi, \Theta) &= \ln \prod_{t \in T} \text{Dir}(\varphi_t; \beta) \prod_{d \in D} \text{Dir}(\theta_d; \alpha) + \text{const} = \\ &= \sum_{t \in T} \sum_{w \in W} (\beta_w - 1) \ln \varphi_{wt} + \sum_{d \in D} \sum_{t \in T} (\alpha_t - 1) \ln \theta_{td}. \end{aligned} \quad (22)$$

Применение уравнений (13)–(14) к этому регуляризатору даёт формулы М-шага:

$$\varphi_{wt} = \text{norm}_{w \in W}(n_{wt} + \beta_w - 1); \quad \theta_{td} = \text{norm}_{t \in T}(n_{td} + \alpha_t - 1).$$

При $\beta_w = 1$, $\alpha_t = 1$ априорное распределение Дирихле совпадает с равномерным распределением на симплексе, формулы М-шага переходят в несмещённые частотные оценки условных вероятностей, а модель LDA переходит в PLSA [45].

При $\beta_w > 1$, $\alpha_t > 1$ регуляризатор имеет сглаживающий эффект: он делает большие вероятности ещё больше, при этом малые вероятности за счёт нормировки становятся меньше, однако никогда не достигают нуля.

При $0 < \beta_w < 1$, $0 < \alpha_t < 1$ регуляризатор имеет разреживающий эффект и способен обнулять малые вероятности.

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, $\{i: p_i > 0\} = \{i: q_i > 0\}$:

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

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 распределения совпадают, $p_i \equiv q_i$.

Если $\text{KL}(P\|Q) < \text{KL}(Q\|P)$, то распределение P сильнее вложено в Q , чем Q в P , см. рис. 3. Таким образом, KL является мерой вложенности двух распределений.

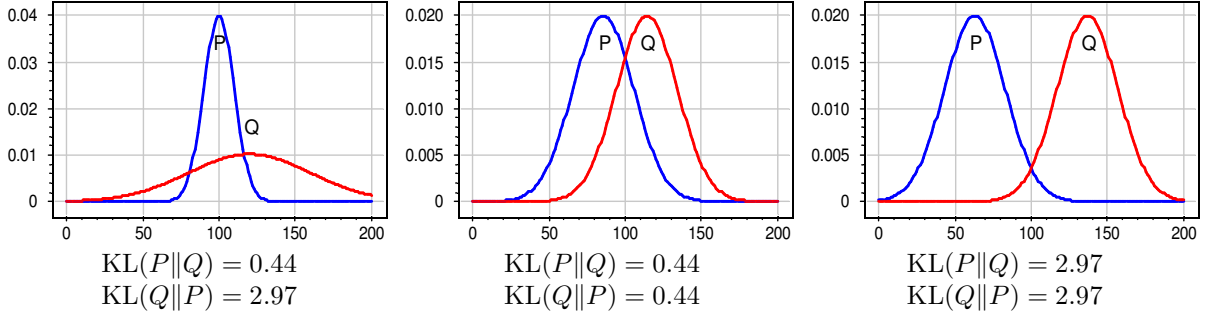


Рис. 3: Дивергенция Кульбака–Лейблера $KL(P||Q)$ является несимметричной мерой вложенности распределения $P = (p_i)_{i=1}^n$ в распределение $Q = (q_i)_{i=1}^n$. Вложенность P в Q приблизительно одинакова на левом и среднем графиках, вложенность Q в P — на левом и правом графиках.

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

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

Максимизация правдоподобия (8) эквивалентна минимизации взвешенной суммы KL-дивергенций между эмпирическими распределениями $\hat{p}(w|d) = \frac{n_{dw}}{n_d}$ и модельными $p(w|d)$, по всем документам d из D :

$$\sum_{d \in D} n_d KL_w \left(\frac{n_{dw}}{n_d} \parallel \sum_{t \in T} \varphi_{wt} \theta_{td} \right) \rightarrow \min_{\Phi, \Theta},$$

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.

Не-вероятностная интерпретация модели LDA. Регуляризатор (22) можно эквивалентным образом записать через KL-дивергенции:

$$R(\Phi, \Theta) = |W| \sum_{t \in T} KL_w \left(\frac{1}{|W|} \parallel \varphi_{wt} \right) - \beta_0 \sum_{t \in T} KL_w \left(\frac{\beta_w}{\beta_0} \parallel \varphi_{wt} \right) + \\ + |T| \sum_{d \in D} KL_t \left(\frac{1}{|T|} \parallel \theta_{td} \right) - \alpha_0 \sum_{d \in D} KL_t \left(\frac{\alpha_t}{\alpha_0} \parallel \theta_{td} \right).$$

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. Дальше всего от равномерного распределения находятся вырожденные распределения, в которых единичная вероятность сконцентрирована в единственном элементе. Поэтому разреживание приводит к обнулению малых вероятностей в матрицах Φ, Θ .

5 Интерпретируемость тем

Отказ от априорных распределений Дирихле позволяет обобщить модель LDA: снять ограничения на знаки гиперпараметров в (22) и свободнее обращаться со сглаживанием и разреживанием для улучшения интерпретируемости тем.

Гипотеза разреженности является одним из естественных необходимых условий интерпретируемости. Предполагается, что каждая тема характеризуется небольшим числом термов, и каждый документ относится к небольшому числу тем. В таком случае среди вероятностей φ_{wt} и θ_{td} должно быть много нулевых значений.

Многочисленные попытки разреживания модели LDA приводили к чрезмерно сложным конструкциям [115, 40, 139, 63, 32] из-за внутреннего противоречия между требованиями разреженности и ограничениями строгой положительности параметров в распределении Дирихле. Проблема решается неожиданно просто, если оставить кросс-энтропийный регуляризатор (22) и разрешить гиперпараметрам α_t, β_w принимать любые значения, включая отрицательные. По всей видимости, впервые она была предложена в динамической модели PLSA для обработки видеопотоков [127], где документами являлись короткие видеофрагменты, термами — признаки на изображениях, темами — появление определённого объекта в течение определённого времени, например, проезд автомобиля. Сильно разреженные распределения потребовались для описания тем с кратким «временем жизни».

Сглаживание и разреживание. По аналогии с (22) введём обобщённый регуляризатор сглаживания и разреживания:

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

Подставив этот регуляризатор в (13)–(14), получим формулы M-шага:

$$\varphi_{wt} = \operatorname{norm}_{w \in W}(n_{wt} + \beta_{wt}); \quad (23)$$

$$\theta_{td} = \operatorname{norm}_{t \in T}(n_{td} + \alpha_{td}). \quad (24)$$

Положительное значение параметра α_{td} или β_{wt} соответствует сглаживанию, отрицательное — разреживанию.

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

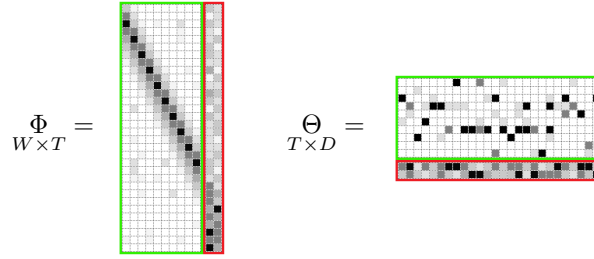


Рис. 4: Структура разреженности матриц Φ и Θ с предметными и фоновыми темами.

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

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

где β_{\pm} и α_{\pm} — коэффициенты регуляризации.

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. 4. 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 [30, 103], 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 [15]. 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» [27]. 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 [53], поиска в социальных медиа информации, связанной с болезнями, симптомами и ме-

тодами лечения [96, 97], с преступностью и экстремизмом [72, 114], с национальностями и межнациональными отношениями [27, 58, 94].

В модели ATAM (ailment topic aspects model) в качестве сглаживающего распределения β_{wt} использовалась большая коллекция медицинских статей [97].

В моделях SSLDA (semi-supervised LDA) и ISLDA (interval semi-supervised LDA) для поиска этно-релевантных тем в постах социальных сетей использовалось сглаживание по словарю из нескольких сотен этнонимов [27]. 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 [16, 17]. 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 (13), имеет вид

$$\varphi_{wt} = \operatorname{norm}_{w \in W} \left(n_{wt} - \tau \varphi_{wt} \sum_{s \in T \setminus t} \varphi_{ws} \right). \quad (25)$$

Этот регуляризатор контрастирует строки матрицы Φ . В каждой строке, независимо от остальных, вероятности φ_{wt} наиболее значимых тем термина w увеличиваются, вероятности остальных тем уменьшаются и могут обращаться в нуль. Разреживание — это сопутствующий эффект декоррелирования. В [123] 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 [6, 134, 133].

Декоррелирование впервые было предложено в модели TWC-LDA (topic-weak-correlated LDA) в рамках байесовского подхода [123]. Соответствующее априорное распределение не является сопряжённым к мультиномиальному, поэтому байесовский вывод сталкивается с техническими трудностями. В ARTM расчётные формулы выводятся в одну строку.

Комбинирование регуляризаторов сглаживания фоновых тем, разреживания предметных тем в матрице Θ и декоррелирования столбцов матрицы Φ использовалось уже во многих работах для улучшения интерпретируемости тем [6, 133, 134,

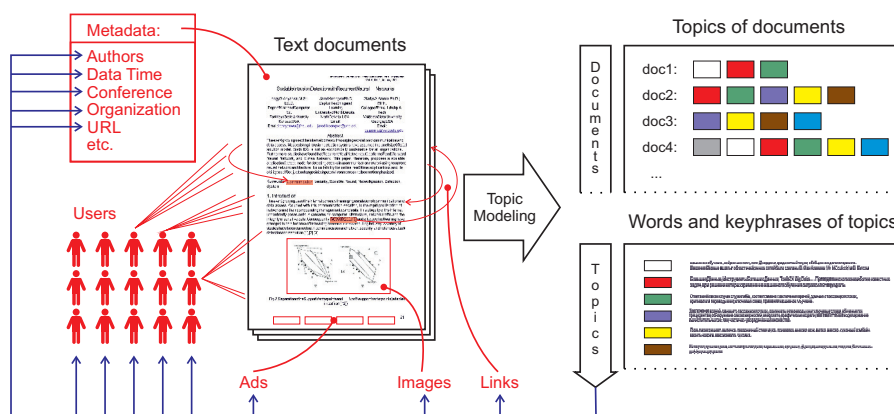


Рис. 5: Обычная тематическая модель определяет распределение тем в каждом документе $p(t|d)$ и распределение термов в каждой теме $p(w|t)$. Мультимодальная модель распространяет семантику тем на элементы всех остальных модальностей, в том числе нетекстовые.

135, 12]. Подобрав коэффициенты регуляризации, можно одновременно значительно улучшить разреженность, контрастность, чистоту и когерентность тем при незначительной потере правдоподобия модели [134]. 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 была использована для тематического разведочного поиска в [12]. Оказалось, что она существенно улучшает качество поиска, хотя никакие критерии качества поиска непосредственно не оптимизировались.

6 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 [138, 147], теги [60], именованные сущности [88] — это примеры текстовых модальностей. В мультязычных тематических моделях параллельных текстов модальностями являются языки [136]. Для анализа коротких текстов с опечатками используют модальность буквенных n -грамм, что позволяет улучшать качество информационного поиска [51]. Примерами нетекстовых модальностей являются (рис. 5): авторы [110], моменты времени [126, 159, 127], классы, жанры или категории [111, 162], цитируемые или цитирующие документы [39] или авторы [56], пользователи электронных библиотек, социальных сетей или рекомендательных систем [64, 118, 140, 155, 156], графические элементы изображений [25, 50, 68], рекламные объявления на веб-страницах [100].

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

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

Тематическая модель модальности m аналогична модели (2):

$$p(w|d) = \sum_{t \in T} p(w|t) p(t|d) = \sum_{t \in T} \varphi_{wt} \theta_{td}, \quad w \in W_m, \quad d \in D. \quad (26)$$

Каждой модальности m соответствует стохастическая матрица $\Phi_m = (\varphi_{wt})_{W_m \times T}$. 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:

$$\sum_{m \in M} \tau_m \sum_{d \in D} \sum_{w \in W_m} n_{dw} \ln \sum_{t \in T} \varphi_{wt} \theta_{td} + R(\Phi, \Theta) \rightarrow \max_{\Phi, \Theta}; \quad (27)$$

$$\sum_{w \in W_m} \varphi_{wt} \in \{0, 1\}; \quad \varphi_{wt} \geq 0; \quad \sum_{t \in T} \theta_{td} \in \{0, 1\}; \quad \theta_{td} \geq 0. \quad (28)$$

Теорема 2. Пусть функция $R(\Phi, \Theta)$ непрерывно дифференцируема. Точка (Φ, Θ) локального экстремума задачи (27)–(28) удовлетворяет системе уравнений со вспомогательными переменными $p_{tdw} = p(t|d, w)$:

$$p_{tdw} = \operatorname{norm}_{t \in T}(\varphi_{wt} \theta_{td}); \quad (29)$$

$$\varphi_{wt} = \operatorname{norm}_{w \in W_m} \left(n_{wt} + \varphi_{wt} \frac{\partial R}{\partial \varphi_{wt}} \right); \quad n_{wt} = \sum_{d \in D} \tau_{m(w)} n_{dw} p_{tdw}; \quad (30)$$

$$\theta_{td} = \operatorname{norm}_{t \in T} \left(n_{td} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right); \quad n_{td} = \sum_{m \in M} \sum_{w \in W_m} \tau_m n_{dw} p_{tdw}. \quad (31)$$

Теорема 1 является частным случаем теоремы 2 в случае, когда модальность только одна, $|M| = 1$ и $\tau_m = 1$. Таким образом, переход от одной модальности к произвольному числу модальностей сводится к двум поправкам:

- 1) матрица Φ разбивается на блоки Φ_m , которые нормируются по-отдельности;
- 2) исходные данные n_{dw} домножаются на веса модальностей $\tau_{m(w)}$.

В проекте **BigARTM** реализована возможность комбинировать любое число модальностей с любыми регуляризаторами [16].

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 [37, 83, 93] 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-

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	забывать	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. 6: Примеры тем из двуязычной тематической модели Википедии. Показаны первые 10 слов каждой темы и их вероятности $p(w|t)$ в процентах.

язычного поиска. Попытки более точного и трудоёмкого выравнивания по предложениям или по словам практически не улучшают качество поиска. Обстоятельный обзор мультязычных тематических моделей можно найти в [136].

На рис. 6 показаны некоторые из 400 тем, построенных по 216 175 парам русских и английских статей Википедии [130]. Для связывания языков использовались только модальности, выравнивания и словари не использовались. Ассессор оценил 396 тем из 400 как хорошо интерпретируемые.

Для использования двуязычного словаря в [7] 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 близки в смысле KL-дивергенции:

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

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

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

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 :

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

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

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 $p(\langle \text{total} \rangle | \langle \text{сумма} \rangle, t) > 0.9$

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

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

$$\varphi_{wt} = \operatorname{norm}_{w \in W^\ell} \left(n_{wt} + \tau \sum_u \pi_{uwt} n_{ut} \right);$$

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

Связывание параллельных текстов сильнее улучшает качество поиска, чем оба способа учёта словарей [7]. 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. 7.

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

$$p(w|d) = \sum_{t \in T} p(w|t) \sum_{c \in C_d} p(t|c) p(c|d) = \sum_{t \in T} \sum_{c \in C_d} \varphi_{wt} \psi_{tc} \pi_{cd}. \quad (32)$$

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:

$$\sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \sum_{c \in C_d} \varphi_{wt} \psi_{tc} \pi_{cd} + R(\Phi, \Psi, \Pi) \rightarrow \max_{\Phi, \Psi, \Pi}; \quad (33)$$

$$\sum_{w \in W} \varphi_{wt} \in \{0, 1\}, \varphi_{wt} \geq 0; \quad \sum_{t \in T} \psi_{tc} \in \{0, 1\}, \psi_{tc} \geq 0; \quad \sum_{c \in C_d} \pi_{cd} \in \{0, 1\}, \pi_{cd} \geq 0. \quad (34)$$

Теорема 3. Пусть функция $R(\Phi, \Psi, \Pi)$ непрерывно дифференцируема. Точка локального экстремума (Φ, Ψ, Π) задачи (33), (34) удовлетворяет системе уравнений со вспомогательными переменными $p_{tcdw} = p(t, c|d, w)$:

$$\begin{aligned} p_{tcdw} &= \operatorname{norm}_{(t,c) \in T \times C_d} (\varphi_{wt} \psi_{tc} \pi_{cd}); \\ \varphi_{wt} &= \operatorname{norm}_{w \in W} \left(n_{wt} + \varphi_{wt} \frac{\partial R}{\partial \varphi_{wt}} \right); & n_{wt} &= \sum_{d \in D} \sum_{c \in C_d} n_{dw} p_{tcdw}; \\ \psi_{tc} &= \operatorname{norm}_{t \in T} \left(n_{tc} + \psi_{tc} \frac{\partial R}{\partial \psi_{tc}} \right); & n_{tc} &= \sum_{d \in D} \sum_{w \in d} n_{dw} p_{tcdw}; \\ \pi_{cd} &= \operatorname{norm}_{c \in C_d} \left(n_{cd} + \pi_{cd} \frac{\partial R}{\partial \pi_{cd}} \right); & n_{cd} &= \sum_{w \in d} \sum_{t \in T} n_{dw} p_{tcdw}. \end{aligned}$$

Данная модель, основанная на трёхматричном разложении, наиболее известна как *автор-тематическая модель* ATM (author-topic model), в которой порождающей модальностью являются авторы документов [110]. В *тематической модели тегирования документов* TWTM (tag weighted topic model) порождающей модальностью являются теги документа [66]. Аналогичная модель использовалась для обработки видеопотоков в [50]. Документы d соответствовали последовательным 1-секундным видеоклипам, термы w — элементарным визуальным событиям, темы t — простым 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 (32) можно упростить и свести снова к двуматричному разложению, если отождествить темы с категориями, $C \equiv T$, и взять единичную матрицу Ψ . Данная модель известна в литературе как Flat-LDA [111] и Labeled-LDA [107]. 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, проследить 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 ТОТ (topics over time) [146] 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:

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

Оба регуляризатора можно записать и как функцию от Φ , и как функцию от Θ . В случае регуляризатора $R_2(\Phi)$ формула М-шага имеет вид²

$$\varphi_{it} = \operatorname{norm}_{i \in I} (n_{it} + \tau_2 \varphi_{it} \operatorname{sign}(\varphi_{i-1,t} - \varphi_{it}) + \tau_2 \varphi_{it} \operatorname{sign}(\varphi_{i+1,t} - \varphi_{it})), \quad (35)$$

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.

²Дойков Н. В. Адаптивная регуляризация вероятностных тематических моделей. Бакалаврская диссертация, ВМК МГУ, 2015.

http://www.MachineLearning.ru/wiki/images/9/9f/2015_417_DoykovNV.pdf

7 Моделирование транзакционных данных

Обычные тематические модели текстовых коллекций описывают вхождения слов в документы. Мультимодальные модели описывают документы, в которых содержатся токены различных модальностей: слова, биграммы, теги, категории, авторы, пользователи, и т. д. Во всех этих случаях модель описывает парные взаимодействия между документами и токенами. В более сложных приложениях исходные данные могут описывать транзакции (отношения, взаимосвязи, взаимодействия) не только между парами объектов различных модальностей, но и между тройками. Например, в сети интернет-рекламы «пользователь u кликнул объявление b на странице d »; в социальной сети «пользователь u написал слово w на странице блога d »; в сети продаж «покупатель b купил у продавца s товар g »; в рекомендательной системе «клиент u оценил фильм f в ситуативном контексте s ».

В этих примерах взаимодействие трёх объектов не сводится к парным взаимодействиям. 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 токенам различных модальностей, рёбра — транзакциям. Задача заключается в том, чтобы по наблюдаемой выборке транзакций восстановить неизвестные тематики вершин $p(t|v)$. Предполагается, что вероятность транзакции тем выше, чем более схожи тематики её вершин.

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

Тематические модели на гиперграфах. Гиперграф $\Gamma = \langle V, E \rangle$ определяется множеством вершин (токенов) V и множеством рёбер (транзакций) E . Каждое ребро $e \in E$ является 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:

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

Каждая транзакция $e \in E$ имеет *тип* $k = \kappa(e)$ из конечного множества типов K . Выборка транзакций разбивается на непересекающиеся подмножества по типам:

$$E = \bigsqcup_{k \in K} E_k, \quad E_k = \{e \in E : \kappa(e) = k\}.$$

Например, в обычных тематических моделях вводятся только две модальности: документы $V_1 = D$ и термы $V_2 = W$; единственным типом транзакций является вхождение термина w в документ d ; каждая транзакция представляется ребром из двух вершин $e = (d, w)$; гиперграф является двудольным графом.

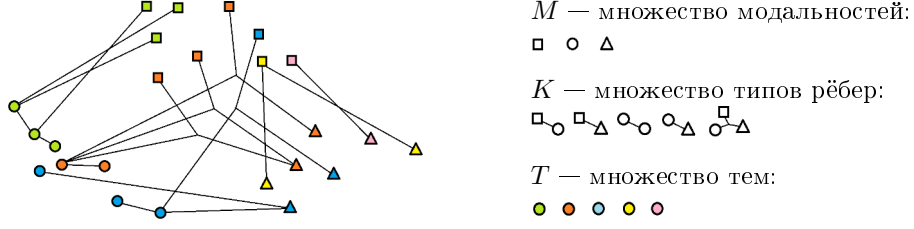


Рис. 8: Пример гиперграфа с вершинам трёх модальностей, рёбрами-транзакциями пяти типов и пятью темами.

На рис. 8 показан пример гиперграфа с вершинами трёх модальностей, рёбрами-транзакциями пяти типов и пятью темами.

Каждому типу рёбер k соответствует дискретное вероятностное пространство $\Omega_k \subseteq 2^V \times T$ с функцией вероятности $p_k: \Omega_k \rightarrow [0, 1]$. Предполагается, что *транзакционные данные* — это объединение $|K|$ выборок независимых наблюдений $(e, t) \in \Omega_k$, каждая из которых порождается своим распределением $p_k(e, t)$, $k \in K$. Каждое ребро гиперграфа $e \in E_k$ входит в выборку n_e раз, и с каждым вхождением ребра связана своя латентная тема $t \in T$.

Будем полагать, что для каждого типа рёбер k первая модальность является *контейнером*, и обозначать через $(d, x) \in E_k$ произвольное ребро типа k , где x — множество всех остальных вершин ребра, за исключением d . Обозначим через D множество всех вершин-контейнеров в гиперграфе.

Сделаем два основных допущения, на которых будет основана вероятностная модель порождения транзакционных данных.

Во-первых, предположим, что распределение тем в вершине-контейнере d не зависит от типа ребра: $p_k(t|d) = p(t|d)$ для всех $k \in K$. Это обобщение обычного для мультимодальных тематических моделей предположения, что распределение тем в документе является общим для всех модальностей.

В то же время, распределения вершин в темах $p_k(v|t)$ не предполагаются одинаковыми для всех типов рёбер k . Например, распределения слов, используемых в текстах веб-страниц, в пользовательских запросах и в рекламных баннерах, могут значимо различаться для одной и той же темы. Дополнительное требование, чтобы эти распределения были похожи, может быть учтено с помощью регуляризатора.

Во-вторых, введём гипотезу условной независимости вершин в рёбрах (d, x) :

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

При сделанных допущениях процесс порождения каждого 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 из своего распределения $p_k(v|t)$ над множеством V_m .

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

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

Параметрами этой модели являются условные вероятности вершин в темах $\varphi_{kvt} = p_k(v|t)$, нормированные по каждой модальности $v \in V_m$, и условные вероятности тем в контейнерах $\theta_{td} = p(t|d)$. Вероятности $p_k(d)$ легко оцениваются по наблюдаемым данным и не зависят от параметров модели.

Таким образом, гиперграфовая тематическая модель определяется:

- 1) гиперграфом $\Gamma = \langle V, E \rangle$,
- 2) множеством модальностей M ,
- 3) разбиением множества вершин по модальностям $\mu: V \rightarrow M$,
- 4) множеством типов рёбер K ,
- 5) разбиением множества рёбер по типам $\kappa: E \rightarrow K$,
- 6) множеством тем T ,
- 7) вероятностным пространством Ω_k с распределением p_k для каждого $k \in K$,
- 8) параметрами модели $\varphi_{kvt} = p_k(v|t)$ и $\theta_{td} = p(t|d)$.

Гиперграфовая модель описывает широкий класс тематических моделей. В частности, она соответствует модели PLSA в случае, когда модальности две — документы $V_1 = D$ и термины $V_2 = W$, тип рёбер только один — пары $(d, w) \in D \times W$.

Для оптимизации параметров модели применим принцип максимума правдоподобия для каждого типа рёбер k и будем максимизировать взвешенную сумму логарифмов правдоподобия с весами τ_k и регуляризатором $R(\Phi, \Theta)$:

$$\sum_{k \in K} \tau_k \sum_{dx \in E_k} n_{dx} \ln \sum_{t \in T} \theta_{td} \prod_{v \in x} \varphi_{kvt} + R(\Phi, \Theta) \rightarrow \max_{\Phi, \Theta}; \quad (36)$$

$$\sum_{v \in V_m} \varphi_{kvt} \in \{0, 1\}, \quad \varphi_{kvt} \geq 0; \quad \sum_{t \in T} \theta_{td} \in \{0, 1\}, \quad \theta_{td} \geq 0.$$

Ограничения-равенства предусматривают возможность обнуления распределений. Если $\varphi_{kvt} = 0$ для всех $v \in V_m$, то тема t не участвует в порождении рёбер типа k с вершинами модальности m . Если $\theta_{td} = 0$ для всех $t \in T$, то считается, что тематическая модель не в состоянии описать содержимое контейнера d .

Теорема 4. Пусть функция $R(\Phi, \Theta)$ непрерывно дифференцируема. Точка локального максимума (Φ, Θ) задачи (36) удовлетворяет системе уравнений относительно параметров модели φ_{kvt} , θ_{td} и вспомогательных переменных $p_{ktdx} = p_k(t|d, x)$:

$$p_{ktdx} = \mathop{\text{norm}}_{t \in T} \left(\theta_{td} \prod_{v \in x} \varphi_{kvt} \right). \quad (37)$$

$$\varphi_{kvt} = \mathop{\text{norm}}_{v \in V_m} \left(n_{kvt} + \varphi_{kvt} \frac{\partial R}{\partial \varphi_{kvt}} \right); \quad n_{kvt} = \sum_{dx \in E_k} [v \in x] \tau_k n_{dx} p_{ktdx}; \quad (38)$$

$$\theta_{td} = \mathop{\text{norm}}_{t \in T} \left(n_{td} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right); \quad n_{td} = \sum_{k \in K} \sum_{dx \in E_k} \tau_k n_{dx} p_{ktdx}. \quad (39)$$

Доказательство. Воспользуемся условиями Каруша–Куна–Таккера. Запишем лагранжиан оптимизационной задачи (36):

$$\begin{aligned} \mathcal{L}(\Phi, \Theta) = & \sum_{k \in K} \tau_k \sum_{dx \in E_k} n_{dx} \ln \sum_{t \in T} \theta_{td} \prod_{v \in x} \varphi_{kvt} + R(\Phi, \Theta) - \\ & - \sum_{k \in K} \sum_{m \in M} \sum_{t \in T} \lambda_{kmt} \left(\sum_{v \in V_m} \varphi_{kvt} - 1 \right) - \sum_{k \in K} \sum_{m \in M} \sum_{v \in V_m} \sum_{t \in T} \lambda_{kmt} \varphi_{kvt} - \\ & - \sum_{d \in D} \mu_d \left(\sum_{t \in T} \theta_{td} - 1 \right) - \sum_{d \in D} \sum_{t \in T} \mu_{td} \theta_{td}. \end{aligned}$$

Приравняем нулю производные лагранжиана по параметрам модели:

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \varphi_{kvt}} = & \sum_{dx \in E_k} [v \in x] \tau_k n_{dx} \frac{\theta_{td} \prod_{u \in x \setminus v} \varphi_{utk}}{p_k(x|d)} + \frac{\partial R}{\partial \varphi_{kvt}} - \lambda_{k\mu(v)t} - \lambda_{k\mu(v)t} \varphi_{kvt} = 0; \\ \frac{\partial \mathcal{L}}{\partial \theta_{td}} = & \sum_{k \in K} \sum_{dx \in E_k} \tau_k n_{dx} \frac{\prod_{v \in x} \varphi_{kvt}}{p_k(x|d)} + \frac{\partial R}{\partial \theta_{td}} - \mu_d - \mu_{td} = 0. \end{aligned}$$

Умножим левую и правую части первого равенства на φ_{kvt} , второго — на θ_{td} :

$$\begin{aligned} \sum_{dx \in E_k} [v \in x] \tau_k n_{dx} \underbrace{\frac{\theta_{td} \prod_{u \in x} \varphi_{utk}}{p_k(x|d)}}_{p_{ktdx} = p_k(t|d,x)} + \varphi_{kvt} \frac{\partial R}{\partial \varphi_{kvt}} = & \lambda_{k\mu(v)t} \varphi_{kvt}; \\ \sum_{k \in K} \sum_{dx \in E_k} \tau_k n_{dx} \underbrace{\frac{\theta_{td} \prod_{v \in x} \varphi_{kvt}}{p_k(x|d)}}_{p_{ktdx} = p_k(t|d,x)} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} = & \mu_d \theta_{td}. \end{aligned}$$

Запишем эти уравнения короче через переменные n_{kvt} из (38) и n_{td} из (39):

$$\varphi_{kvt} \lambda_{k\mu(v)t} = n_{kvt} + \varphi_{kvt} \frac{\partial R}{\partial \varphi_{kvt}}; \quad (40)$$

$$\theta_{td} \mu_d = n_{td} + \theta_{td} \frac{\partial R}{\partial \theta_{td}}. \quad (41)$$

Если $\lambda_{kmt} \leq 0$, то тема t вырождена: $\varphi_{kvt} = 0$ для всех $v \in V_m$.

Если $\lambda_{kmt} > 0$, то либо обе части равенства (40) положительны, либо $\varphi_{kvt} = 0$. Объединяя оба случая в одну формулу, получим

$$\varphi_{kvt} \lambda_{k\mu(v)t} = \left(n_{kvt} + \varphi_{kvt} \frac{\partial R}{\partial \varphi_{kvt}} \right)_+. \quad (42)$$

Если $\mu_d \leq 0$, то контейнер d вырожден: $\theta_{td} = 0$ для всех $t \in T$.

Если $\mu_d > 0$, то либо обе части равенства (41) положительны, либо $\theta_{td} = 0$. Объединяя оба случая в одну формулу, получим

$$\theta_{td} \mu_d = \left(n_{td} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right)_+. \quad (43)$$

Просуммируем равенства (42) по $v \in V_m$, равенства (43) по $t \in T$, применим условия нормировки и выразим двойственные переменные:

$$\begin{aligned} \lambda_{kmt} = & \sum_{v \in V_m} \left(n_{kvt} + \varphi_{kvt} \frac{\partial R}{\partial \varphi_{kvt}} \right)_+; \\ \mu_d = & \sum_{t \in T} \left(n_{td} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right)_+. \end{aligned}$$

Подставив λ_{kmt} и μ_d в (42) и (43), получим (38) и (39). Теорема доказана.

8 Моделирование зависимостей

Классификация. Тематическая модель классификации Dependency LDA [111] является байесовским аналогом модели (26) 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 $p(c|d) \geq \gamma_c$.

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 [111] показали, что тематические модели превосходят обычные методы многоклассовой классификации на больших текстовых коллекциях с большим числом несбалансированных, пересекающихся, взаимозависимых классов. В [131] 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 $C'_d \subset C$. Для этого случая запишем правдоподобие вероятностной модели бинарных данных:

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

Первое слагаемое есть log-правдоподобие модальности классов (26), 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 [73, 111]. 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 $D_c = \{d \in D : c \in C_d\}$ — множество документов, относящихся к классу c . Выразим данное требование с помощью сглаживающего регуляризатора кросс-энтропии, который можно интерпретировать и как максимизацию правдоподобия модели классов $p(c)$:

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

где $n_t = \sum_c n_{ct}$ — число токенов модальности C , относящихся к теме t во всей коллекции. Подставляя этот регуляризатор в (13), получим формулы М-шага:

$$\varphi_{ct} = \operatorname{norm}_{w \in W} \left(n_{ct} + \tau |D_c| \frac{n_t \varphi_{ct}}{\sum_s n_s \varphi_{cs}} \right). \quad (44)$$

Частотная регуляризация использовалась в тематической модели Prior-LDA, которая была предложена в [111] 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 [77, 120].

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:

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

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

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

Подставляя этот регуляризатор в (14) и приравнивая нулю его производную по v , получим формулы М-шага:

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

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

Заметим, что 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 онлайн-овом EM-алгоритме.

В [120] показано, что качество регрессии может зависеть от инициализации тематической модели, и предложено несколько методов инициализации.

Корреляции тем. *Модель коррелированных тем* СТМ (correlated topic model) предназначена для выявления связей между темами [21]. 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 вектор-столбцов 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 :

$$p(\eta_d | \mu, \Sigma) = (2\pi)^{-\frac{n}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\eta_d - \mu)^\top \Sigma^{-1} (\eta_d - \mu)\right).$$

Изначально модель СТМ 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 СТМ формализуется и реализуется намного проще.

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

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

Согласно (14), формула M-шага для θ_{td} принимает вид

$$\theta_{td} = \text{norm}_{t \in T} \left(n_{td} - \tau \sum_{s \in T} \Sigma_{ts}^{-1} (\ln \theta_{sd} - \mu_s) \right), \quad (45)$$

где Σ_{ts}^{-1} — элементы обратной ковариационной матрицы. Параметры Σ, μ нормального распределения обновляются после каждого прохода коллекции, либо после каждого пакета документов в онлайн-овом EM-алгоритме:

$$\begin{aligned} \mu &= \frac{1}{|D|} \sum_{d \in D} \ln \theta_d; \\ \Sigma &= \frac{1}{|D|} \sum_{d \in D} (\ln \theta_d - \mu) (\ln \theta_d - \mu)^\top. \end{aligned}$$

Таким образом, трудоёмкая операция обращения ковариационной матрицы выполняется относительно редко. В [21] использовалась LASSO-регрессия, чтобы получать разреженную ковариационную матрицу.

9 Определение числа тем

Регуляризатор отбора тем предложен в [133] для удаления незначимых тем из тематической модели. Он основан на идее кросс-энтропийного разреживания распределения $p(t)$, которое легко выражается через параметры тематической модели:

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

Подставим этот регуляризатор в формулу М-шага (14):

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

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

$$\theta_{td} = \operatorname{norm}_{t \in T} \left(n_{td} \left(1 - \tau \frac{n}{n_t |T|} \right) \right). \quad (46)$$

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 постепенно увеличивается.

Отбор тем в ARTM намного проще непараметрических байесовских моделей — иерархического процесса Дирихле (hierarchical Dirichlet process, HDP) [124] или процесса китайского ресторана (Chinese restaurant process, CRP) [24].

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 [135] 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. При этом ARTM определяет его более точно и устойчиво. 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. В реальных текстовых коллекциях никакого «истинного числа тем» просто не существует. Чем больше коллекция, тем более мелкие семантические различия, и для большего числа тем, возможно уловить.

Эти соображения подтверждаются опытом построения иерархических тематических моделей. Темы можно дробить на более мелкие подтемы вплоть до порога статистической значимости. Выбор этого порога также является эвристикой, и от него зависит итоговое число тем.

В ходе экспериментов [135] также выяснилось, что регуляризатор отбора тем имеет полезный сопутствующий эффект: он удаляет из модели дублирующие, расщеплённые и линейно зависимые темы.

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

10 Построение тематических иерархий

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 [23, 67, 82, 158, 105, 141, 142, 143, 144], оптимизация размера и структуры иерархии остаётся открытой проблемой; более того, оценивание качества иерархий — также открытая проблема [158].

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.

В [33] 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 в темах. В мультидисциплинарных коллекциях подтемы могут иметь по несколько родительских тем.

Регуляризатор межуровневых связей. 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 :

$$\sum_{t \in T} n_t \text{KL}_w \left(p(w|t) \parallel \sum_{s \in S} p(w|s) p(s|t) \right) = \sum_{t \in T} n_t \text{KL}_w \left(\frac{n_{wt}}{n_t} \parallel \sum_{s \in S} \varphi_{ws} \psi_{st} \right) \rightarrow \min_{\Phi, \Psi},$$

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} :

$$R(\Phi, \Psi) = \tau \sum_{t \in T} \sum_{w \in W} n_{wt} \ln \sum_{s \in S} \varphi_{ws} \psi_{st}. \quad (47)$$

Задача максимизации $R(\Phi, \Psi)$ с точностью до обозначений совпадает с основной задачей тематического моделирования (8), если считать родительские темы t псевдодокументами с частотами термов $\tau n_{wt} = \tau n_t \varphi_{wt}$. Это означает, что вместо добавления слагаемого в формулы М-шага данный регуляризатор можно реализовать ещё проще. Построив родительский уровень, надо добавить в коллекцию ровно

$|T|$ псевдодокументов, задав им в качестве частот термов значения τn_{wt} . Матрица Ψ получится в столбцах матрицы Θ , соответствующих псевдодокументам [33].

В библиотеке `BigARTM` этот подход реализован в виде отдельного класса `hARTM`.

Регуляризатор разреживания 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 (14), из которой следует формула М-шага для модели дочернего уровня:

$$\psi_{st} = \operatorname{norm}_{s \in S} \left(n_{st} + \tau \left(p(t|s) - \frac{1}{|T|} \right) \right). \quad (48)$$

Согласно этой формуле, условные вероятности $p(t|s)$, меньшие $\frac{1}{|T|}$, становятся ещё меньше, и при достаточно большом τ обнуляются [33].

11 Моделирование связей между документами

Ссылки и цитирование. Иногда имеется дополнительная информация о связях между документами и предполагается, что связанные документы имеют схожую тематику. Связь может означать, что два документа относятся к одной рубрике, совместно упоминаются или ссылаются друг на друга. Формализуем это предположение с помощью регуляризатора:

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

где n_{dc} — вес связи между документами, например, число ссылок из d на c . В [39] 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 (14), имеет вид

$$\theta_{td} = \operatorname{norm}_{t \in T} \left(n_{td} + \tau \theta_{td} \sum_{c \in D} n_{dc} \theta_{tc} \right).$$

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 [39]. 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 [157] определяются регионы популярности национальной кухни по постам пользователей Flickr. Другая иллюстрация из [78] показывает, что тематическая модель, учитывающая, из какого штата США пришло сообщение, точнее прослеживает путь урагана «Катрина».

Квадратичный регуляризатор матрицы Θ , предложенный в [157], формализует предположение, что документы со схожими геолокациями имеют схожую тематику:

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

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:

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

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 [78] 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:

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

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 токен 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 :

$$R(\Phi) = -\frac{\tau}{2} \sum_{(u,v) \in E} w_{uv} \sum_{t \in T} n_t^2 \left(\frac{\varphi_{vt}}{|D_v|} - \frac{\varphi_{ut}}{|D_u|} \right)^2. \quad (49)$$

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 [122] предполагается, что если $(u, v) \in E$, то тематика $p(t|u)$ шире тематики $p(t|v)$. Поэтому минимизируется сумма дивергенций $\text{KL}(p(t|v) \| p(t|u))$, 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 эксперименты³, регуляризация матрицы Φ приводит практически к тем же результатам, что и регуляризация Θ для моделей NetPLSA и iTopicModels.

³Булатов В. Г. Использование графовой структуры в тематическом моделировании. Магистерская диссертация, МФТИ, 2016.

<http://www.MachineLearning.ru/wiki/images/4/4d/Bulatov-2016-ms.pdf>

12 Моделирование совстречаемости слов

Гипотеза «мешка слов» является одним из самых критикуемых постулатов тематического моделирования. Поэтому многие исследования направлены на создание более адекватных моделей, учитывающих порядок слов. Из них наиболее важными представляются три направления.

Первое направление связано с выделением *коллокаций* — статистически устойчивых *n*-грамм (последовательностей подряд идущих *n* слов). Темы, построенные на *n*-граммах, намного лучше интерпретируются, чем построенные на униграммах (отдельных словах). Проблема в том, что число *n*-грамм катастрофически быстро растёт с ростом объёма коллекции.

Второе направление связано с анализом совместной встречаемости слов. Появление программы `word2vec` [79] 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, и в то же время сохраняющие 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 13.

Коллокации. Использование словосочетаний заметно улучшает интерпретируемость тем, что демонстрируется практически в каждой публикации по *n*-граммным тематическим моделям, см. например [54].

Первая биграммная тематическая модель ВТМ (bigram topic model) [138] 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:

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

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

Модель TNG (topical *n*-grams) [147] устраняет эти недостатки. Условное распределение слов описывается вероятностной смесью $p(w|v, t) = \xi_{vwt} \varphi_{wt}^v + (1 - \xi_{vwt}) \varphi_{wt}$, где ξ_{vwt} — переменная, равная вероятности того, что пара слов « vw » является биграммой в теме t . В работе С. С. Стенина⁴ 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

⁴С. С. Стенин. Мультиграммные аддитивно регуляризованные тематические модели. Магистерская диссертация, ФИВТ МФТИ, 2015.

<http://www.MachineLearning.ru/wiki/images/4/4a/Stenin2015MasterThesis.pdf>

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	тупиговое множество	комитет	задача MASC

Рис. 9: Примеры тем униграммной модели и соответствующих им тем биграммной модели (по коллекции статей научной конференции ММРО).

словами, мультимодальная ARTM может быть использована для поиска приближённого решения в модели TNG.

В той же работе были проведены эксперименты с биграммной мультимодальной моделью ARTM на небольшой (менее 1000 документов) коллекции русскоязычных статей научной конференции ММРО (математические методы распознавания образов). Сопоставление тем униграммной и биграммной моделей показало, что по темам биграммной модели опрошенные постоянные участники конференции могли определить научную группу и даже авторов статей, тогда как по темам униграммной модели сделать это было проблематично, см. рис. 9.

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 [41]. 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 [69] и AutoPhrase [113], демонстрирующие ещё лучшие результаты.

Битермы. *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 записях диалогов контакт-центра, 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* (bitern topic model, BTM) [151].

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 бигерма могут находиться другие слова. Конкретизация понятия «рядом» зависит от постановки задачи и особенностей коллекции.

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

Примем гипотезу условной независимости $p(u, v | t) = p(u | t) p(v | t)$, то есть допустим, что слова бигермов 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 $P = \Phi \Pi \Phi^T$, где $\Pi = \operatorname{diag}(\pi_1, \dots, \pi_T)$ — диагональная матрица. Модель бигермов не определяет тематику документов Θ и поэтому не подвержена влиянию эффектов, вызванных короткими текстами.

ARTM позволяет объединить модель бигермов с обычной тематической моделью, чтобы всё-таки получить матрицу Θ . Для этого возьмём \log -правдоподобие модели бигермов 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 (13)–(14) к этому регуляризатору даёт формулы М-шага:

$$\varphi_{wt} = \operatorname{norm}_{w \in W} \left(n_{wt} + \tau \sum_{u \in W} n_{uw} p_{tuw} \right); \quad (50)$$

$$p_{tuw} = \operatorname{norm}_{t \in T} (n_t \varphi_{wt} \varphi_{ut}). \quad (51)$$

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 $\theta_{tu} = \operatorname{norm}_t (n_t \varphi_{ut})$. Другими словами, в модели бигермов 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 бигермам. В таком случае модель ARTM переходит в модель бигермов, которая строится по коллекции псевдо-документов, без использования исходных документов.

Сеть слов. Идея моделировать не документы, а связи между словами, была положена в основу тематических моделей встречаемости слов WTM (word topic model) [31] и WNTM (word network topic model) [163]. Любопытно, что более ранняя публикация модели WTM осталась незамеченной (видимо, как не-байесовская), и во второй статье даже нет ссылки 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 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} — совстречаемость слов u, w (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 $\Theta = \text{diag}(\pi_1, \dots, \pi_t) \Phi^\top$ и количество параметров вдвое меньше. Как показали эксперименты на коллекциях коротких текстов, модель WNTM немного превосходит модель битермов и существенно превосходит обычные тематические модели [163]. На коллекциях длинных документов тематические модели совстречаемости 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 [90]. Совстречаемость термов может оцениваться по самой коллекции D [84], или по сторонней коллекции, например, по Википедии [87]. Средняя когерентность тем считается хорошей мерой интерпретируемости тематической модели [91].

Пусть заданы оценки совместной встречаемости $C_{wv} = \hat{p}(w | v)$ 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 M-шага, согласно (13), принимает вид

$$\varphi_{wt} = \underset{w \in W}{\text{norm}} \left(n_{wt} + \tau \sum_{v \in W \setminus w} C_{wv} n_{vt} \right). \quad (52)$$

Это сглаживающий регуляризатор. Он увеличивает вероятность терма в теме, если термы, с которыми он часто совместно встречается, относятся к данной теме. Точно такая же формула получилась в [84] 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 [87] 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 совместной встречаемости $C_{uv} = N_{uv}[\text{PMI}(u, v) > 0]$ определяется через *поточечную взаимную информацию* (pointwise mutual information)

$$\text{PMI}(u, v) = \ln \frac{|D|N_{uv}}{N_u N_v}, \quad (53)$$

где N_{uv} — число документов, в которых термины 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 [47]. Слова, встречающиеся в схожих контекстах, имеют схожую семантику и, соответственно, должны иметь близкие векторы. Для формализации этого принципа в [79, 80] 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:

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

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 лог-правдоподобия, как правило, градиентными методами:

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

Постановка задачи очень похожа на тематические модели ВТМ и WNTM. Модели семейства `word2vec` и другие модели векторных представлений слов также являются матричными разложениями [65, 99, 70]. 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 [102] предложен способ построения *тематических векторных представлений слов* (probabilistic word embedding, PWE) по псевдо-коллекции документов, аналогичный моделям ВТМ и WNTM. В задачах семантической близости слов они конкурируют с моделями `word2vec` и существенно превосходят обычные тематические

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

Рис. 10: Примеры решения задач ассоциаций слов для моделей ARTM и word2vec. Приводятся четыре наиболее близкие ассоциации.

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. 10 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 [35], специально разработанную для поиска семантически близких документов.

Кроме того, ARTM позволяет обобщить тематические модели дистрибутивной семантики для мультимодальных коллекций [102]. Используя данные о встречаемости токенов 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.

13 Тематическая сегментация текстов

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 связном тексте. 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 $\varphi_{wt}, \theta_{td}$:

$$p(s|d) = \sum_{t \in T} p(t|d) \prod_{w \in s} p(w|t)^{n_{sw}} = \sum_{t \in T} \theta_{td} \prod_{w \in s} \varphi_{wt}^{n_{sw}}.$$

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 :

$$\sum_{d \in D} \sum_{s \in S_d} \ln \sum_{t \in T} \theta_{td} \prod_{w \in s} \varphi_{wt}^{n_{sw}} + R(\Phi, \Theta) \rightarrow \max_{\Phi, \Theta}, \quad (54)$$

при обычных ограничениях (11). В частном случае, когда каждый сегмент состоит только из одного слова, данная задача переходит в (10). Заметим также, что (54) является частным случаем гиперграфовой модели (36), в которой вершины являются словами, а рёбрами — предложения.

Теорема 5. Пусть функция $R(\Phi, \Theta)$ непрерывно дифференцируема. Точка (Φ, Θ) локального экстремума задачи (54), (11) удовлетворяет системе уравнений со вспомогательными переменными $p_{tds} = p(t|d, s)$:

$$\begin{aligned} p_{tds} &= \operatorname{norm}_{t \in T} \left(\theta_{td} \prod_{w \in s} \varphi_{wt}^{n_{sw}} \right); \\ \varphi_{wt} &= \operatorname{norm}_{w \in W} \left(n_{wt} + \varphi_{wt} \frac{\partial R}{\partial \varphi_{wt}} \right); & n_{wt} &= \sum_{d \in D} \sum_{s \in S_d} [w \in s] p_{tds}; \\ \theta_{td} &= \operatorname{norm}_{t \in T} \left(n_{td} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right); & n_{td} &= \sum_{s \in S_d} p_{tds}. \end{aligned}$$

Аналогичным образом задача ставится для модели предложений senLDA [19] и для модели коротких сообщений Twitter-LDA [160]. В обоих случаях регуляризатором являются априорные распределения Дирихле. В модели Twitter-LDA в роли документов выступают авторы, в роли сегментов — сообщения данного автора.

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* [108] основан на пост-обработке распределений $p(t|d, w_i)$, $i = 1, \dots, n$, получаемых какой-либо тематической моделью, например, LDA. Определим тематику предложения s как среднюю тематику $p(t|d, w)$ всех его слов⁵. Посчитаем 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

⁵Точнее, в [108] предлагалось для каждого слова выбирать наиболее вероятную тему. Оба варианта имеют право на существование. Какой из них лучше, пока не исследовано.

стоп-слов, фоновых тем и коротких предложений. Аккуратная настройка параметров этих эвристик позволяет достичь высокого качества сегментации [108].

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 через распределения $p_{tdw} = p(t|d, w)$, а не через φ_{wt} и θ_{td} . Например, требования сходства тематики термов внутри предложений или между соседними предложениями. Они позволяют учитывать порядок слов внутри документов в обход гипотезы «мешка слов». В общем случае они выражаются в виде регуляризатора $R(\Pi)$, где $\Pi = (p_{tdw})_{T \times D \times W}$ — трёхмерная матрица вспомогательных переменных.

Будем предполагать, что R является достаточно гладкой функцией всех переменных p_{tdw} . Кроме того, сделаем вполне естественное допущение, что если слова w нет в документе d , то функция R не зависит от переменной p_{tdw} .

Согласно уравнению (12), матрица Π является функцией от Φ и Θ . Поэтому к регуляризатору $R(\Pi(\Phi, \Theta))$ в полной мере применима теорема 1. При этом систему уравнений удобнее записать через частные производные регуляризатора по вспомогательным переменным Π .

Рассмотрим задачу максимизации регуляризованного log-правдоподобия с двумя регуляризаторами, один из которых зависит от Π :

$$\sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \varphi_{wt} \theta_{td} + R(\Pi(\Phi, \Theta)) + R'(\Phi, \Theta) \rightarrow \max_{\Phi, \Theta}, \quad (55)$$

при ограничениях неотрицательности и нормировки (11).

Теорема 6. Пусть функции $R(\Pi(\Phi, \Theta))$ и $R'(\Phi, \Theta)$ непрерывно дифференцируемы и функция $R(\Pi)$ не зависит от переменных p_{tdw} в случае $n_{dw} = 0$. Тогда точка (Φ, Θ) локального экстремума задачи (55), (11) удовлетворяет системе уравнений со вспомогательными переменными $p_{tdw} = p(t|d, w)$ и \tilde{p}_{tdw} :

$$p_{tdw} = \operatorname{norm}_{t \in T}(\varphi_{wt} \theta_{td}); \quad (56)$$

$$\tilde{p}_{tdw} = p_{tdw} \left(1 + \frac{1}{n_{dw}} \left(\frac{\partial R(\Pi)}{\partial p_{tdw}} - \sum_{z \in T} p_{zdw} \frac{\partial R(\Pi)}{\partial p_{zdw}} \right) \right); \quad (57)$$

$$\varphi_{wt} = \operatorname{norm}_{w \in W} \left(\sum_{d \in D} n_{dw} \tilde{p}_{tdw} + \varphi_{wt} \frac{\partial R'}{\partial \varphi_{wt}} \right); \quad (58)$$

$$\theta_{td} = \operatorname{norm}_{t \in T} \left(\sum_{w \in d} n_{dw} \tilde{p}_{tdw} + \theta_{td} \frac{\partial R'}{\partial \theta_{td}} \right). \quad (59)$$

Доказательство теоремы удобно оформить в виде следующих трёх лемм.

Лемма 1. Для функции $p_{tdw}(\Phi, \Theta) = \frac{\varphi_{wt} \theta_{td}}{\sum_z \varphi_{wz} \theta_{zd}}$ и любого $z \in T$

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

Доказательство. Воспользуемся формулами (12) для переменных p_{tdw} :

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

Лемма доказана.

Введём функцию от вспомогательных переменных Π :

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

Лемма 2. Пусть $R(\Pi)$ не зависит от переменных p_{tdw} при $w \notin d$. Тогда все частные производные регуляризатора $R(\Pi)$ выражаются через Q_{tdw} :

$$\varphi_{wt} \frac{\partial R(\Pi)}{\partial \varphi_{wt}} = \sum_{d \in D} p_{tdw} Q_{tdw}(\Pi); \quad \theta_{td} \frac{\partial R(\Pi)}{\partial \theta_{td}} = \sum_{w \in d} p_{tdw} Q_{tdw}(\Pi).$$

Доказательство. Воспользуемся формулой дифференцирования сложной функции и тем, что $\frac{\partial p_{zdv}}{\partial \varphi_{wt}} = 0$ при $w \neq v$; $\frac{\partial p_{zdv}}{\partial \theta_{td}} = 0$ при $t \neq z$; $\frac{\partial R(\Pi)}{\partial p_{tdw}} = 0$ при $w \notin d$:

$$\varphi_{wt} \frac{\partial R(\Pi)}{\partial \varphi_{wt}} = \varphi_{wt} \sum_{(z,d,v)} \frac{\partial R(\Pi)}{\partial p_{zdv}} \frac{\partial p_{zdv}}{\partial \varphi_{wt}} = \sum_{d \in D} \sum_{z \in T} \frac{\partial R(\Pi)}{\partial p_{zdw}} \varphi_{wt} \frac{\partial p_{zdw}}{\partial \varphi_{wt}}; \quad (60)$$

$$\theta_{td} \frac{\partial R(\Pi)}{\partial \theta_{td}} = \theta_{td} \sum_{(z,d,v)} \frac{\partial R(\Pi)}{\partial p_{zdv}} \frac{\partial p_{zdv}}{\partial \theta_{td}} = \sum_{w \in d} \sum_{z \in T} \frac{\partial R(\Pi)}{\partial p_{zdw}} \theta_{td} \frac{\partial p_{zdw}}{\partial \theta_{td}}. \quad (61)$$

В силу леммы 1 справедливо тождество

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

Подстановка полученного выражения в (60) и (61) завершает доказательство.

Лемма 3. Решение Φ, Θ задачи (55) удовлетворяет следующей системе уравнений относительно переменных $\varphi_{wt}, \theta_{td}$ и вспомогательных переменных p_{tdw} :

$$p_{tdw} = \operatorname{norm}_{t \in T} (\varphi_{wt} \theta_{td}); \quad (62)$$

$$\varphi_{wt} = \operatorname{norm}_{w \in W} \left(\sum_{d \in D} n_{dw} p_{tdw} + \sum_{d \in D} Q_{tdw} p_{tdw} + \varphi_{wt} \frac{\partial R'}{\partial \varphi_{wt}} \right); \quad (63)$$

$$\theta_{td} = \operatorname{norm}_{t \in T} \left(\sum_{w \in d} n_{dw} p_{tdw} + \sum_{w \in d} Q_{tdw} p_{tdw} + \theta_{td} \frac{\partial R'}{\partial \theta_{td}} \right). \quad (64)$$

Эта лемма является непосредственным следствием теоремы 1 и леммы 2.

Осталось только заметить, что выделение вспомогательной переменной \tilde{p}_{tdw} согласно (57) позволяет переписать уравнения (63) и (64) в требуемом виде (58)–(59).

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 (13)–(14) 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:

$$\text{KL}\left(\frac{1}{|T|} \parallel p(t|d, w)\right) \rightarrow \max.$$

Суммируя по всем термам всех документов, получим регуляризатор:

$$R(\Pi) = -\frac{\tau}{|T|} \sum_{d \in D} \sum_{w \in d} n_{dw} \sum_{t \in T} \ln p_{tdw} \rightarrow \max.$$

Подставим производную

$$\frac{\partial R(\Pi)}{\partial p_{zdw}} = -\frac{\tau}{|T|} \frac{n_{dw}}{p_{zdw}}$$

в формулу (57):

$$\tilde{p}_{tdw} = p_{tdw} - \tau \left(\frac{1}{|T|} - p_{tdw} \right).$$

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 (63)–(64):

$$\begin{aligned} \varphi_{wt} &= \text{norm}_{w \in W} \left(n_{wt} - \tau n_w \left(\frac{1}{|T|} - \frac{n_{wt}}{n_w} \right) + \varphi_{wt} \frac{\partial R'}{\partial \varphi_{wt}} \right); \\ \theta_{td} &= \text{norm}_{t \in T} \left(n_{td} - \tau n_d \left(\frac{1}{|T|} - \frac{n_{td}}{n_d} \right) + \theta_{td} \frac{\partial R'}{\partial \theta_{td}} \right). \end{aligned}$$

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. Важнейшая особенность регуляризатора E-шага заключается в том, что он позволяет учитывать соседство слов внутри документа в обход гипотезы «мешка слов». Применим регуляризацию E-шага для построения тематической модели сегментированного текста. 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:

$$R(\Pi) = -\tau \sum_{d \in D} \sum_{s \in S_d} \sum_{t \in T} \ln \sum_{w \in s} n_{sw} p_{tdw}. \quad (65)$$

Опуская рутинные выкладки, приведём результат подстановки (65) в (57):

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

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.

14 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 ассессоров.

Внешние критерии крайне разнообразны и зависят от решаемой прикладной задачи. Практически в каждой публикации по тематическому моделированию используется какой-либо внешний критерий: качество классификации документов [111], точность и полнота информационного поиска [154, 14, 7, 12], число найденных хорошо интерпретируемых тем [17], качество сегментации текстов [108]. В [34] предлагается методика диагностики моделей, основанная на сопоставлении найденных тем с заранее известными концептами.

Перплексия. Наиболее распространённым внутренним критерием является *перплексия* (perplexity), используемая для оценивания моделей языка в компьютерной лингвистике. Это мера несоответствия или «удивлённости» модели $p(w|d)$ токенам w , наблюдаемым в документах d коллекции D . Она определяется через log-правдоподобие (8), либо через log-правдоподобие (27) каждой модальности m :

$$\text{perp}_m(D; p) = \exp\left(-\frac{1}{n_m} \sum_{d \in D} \sum_{w \in W^m} n_{dw} \ln p(w|d)\right), \quad (66)$$

где $n_m = \sum_{d \in D} \sum_{w \in W^m} n_{dw}$ — длина коллекции по m -й модальности. Чем меньше перплексия, тем лучше модель p предсказывает появление токенов 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 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 ряда посторонних факторов — длины документов, мощности и разреженности словаря. 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 $\text{perp}_m(D; p_D)$ является оптимистично смещённой (заниженной) характеристикой качества модели из-за эффекта переобучения. Обобщающую способность тематических моделей принято оценивать *перплексией контрольной выборки* (hold-out perplexity) $\text{perp}_m(D'; p_D)$. Обычно коллекцию разделяют на обучающую и контрольную случайным образом в пропорции 9 : 1 [26].

Недостатком контрольной перплексии является высокая чувствительность к редким и новым словам, которые практически бесполезны для тематических моделей. В ранних экспериментах было показано, что LDA существенно превосходит PLSA по перплексии, откуда был сделан вывод, что LDA меньше переобучается [26]. В [4, 103, 5] были предложены *робастные тематические модели*, описывающие редкие слова специальным «фоновым» распределением. Перплексия робастных вариантов PLSA и LDA оказалась существенно меньшей и практически одинаковой.

Когерентность. 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 [29]. Свойство интерпретируемости важно в информационно-поисковых системах для систематизации и визуализации результатов тематического поиска или категоризации документов.

Большинство существующих методов оценивания интерпретируемости основано на привлечении экспертов-ассессоров. В [89] экспертам предлагалось непосредственно оценивать полезность тем по трёхбалльной шкале. В методе интрузий [29] для каждой найденной темы составляется список из 10 наиболее частотных слов, в который

внедряется одно случайное слово. Тема считается интерпретируемой, если подавляющее большинство экспертов правильно указывают лишнее слово. Экспертные подходы необходимы на стадии исследований, но они затрудняют автоматическое построение тематических моделей. В серии работ [89, 90, 90, 84] показано, что среди величин, вычисляемых по коллекции автоматически, лучше всего коррелирует с экспертными оценками интерпретируемости *когерентность* (coherence).

Тема называется *когерентной* (согласованной), если термы, наиболее частые в данной теме, неслучайно часто совместно встречаются рядом в документах коллекции [90, 91]. Численной мерой когерентности темы t является поточечная взаимная информация (53), вычисляемая по k наиболее вероятным словам темы:

$$\text{coher}_t = \frac{2}{k(k-1)} \sum_{i=1}^{k-1} \sum_{j=i}^k \text{PMI}(w_i, w_j), \quad (67)$$

где w_i — i -й терм в порядке убывания φ_{wt} , число k обычно полагается равным 10.

Когерентность модели определяется как средняя когерентность coher_t всех тем. Когерентность может оцениваться по сторонней коллекции (например, по Википедии) [87], либо по той же коллекции, по которой строится модель [84].

Разреженность и различность 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 частям матриц Φ, Θ , соответствующим предметным темам.

В [133] вводятся косвенные меры интерпретируемости тем, не требующие привлечения ассессоров. Предполагается, что интерпретируемая тема должна содержать *лексическое ядро* — множество слов, которые с большой вероятностью употребляются в данной теме и редко употребляются в других темах. В таком случае матрицы Φ и Θ должны обладать структурой разреженности, аналогичной рис. 4.

Ядро $W_t = \{w \in W \mid p(t|w) > 0.25\}$ темы t определяется как множество термов, которые имеют высокую условную вероятность $p(t|w) = \varphi_{wt} \frac{n_t}{n_w}$ для данной темы. Затем по ядру определяется три показателя интерпретируемости темы t :

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

Показатели размера ядра, чистоты и контрастности для модели в целом определяются как средние по всем предметным темам $t \in S$.

Доля фоновых слов во всей коллекции принимает значения от 0 до 1:

$$\text{BackRatio} = \frac{1}{n} \sum_{d \in D} \sum_{w \in d} \sum_{t \in B} n_{dw} p(t|d, w).$$

Если её значение близко к 0, то модель не способна выделять слова общей лексики, если же оно близко к 1, то это свидетельствует о вырождении тематической модели.

Такие критерии, как размер ядра или доля фоновых слов, могут использоваться для контроля адекватности модели. Чрезмерная регуляризация может приводить

к деградации тем или вырождению модели для слишком большой доли документов. Образно говоря, регуляризаторы в малых дозах являются лекарствами, но в случае передозировки могут превращаться в яд. Поэтому нужны критерии, чувствительны к различным типам вырождения модели.

На практике к тематическим моделям предъявляются сочетания разнообразных требований, для построения модели применяется многокритериальная оптимизация, поэтому и качество модели должно оцениваться по многим критериям.

В проекте BigARTM поддерживается библиотека стандартных метрик качества и механизмы добавления новых пользовательских метрик.

15 Проект BigARTM

BigARTM — это библиотека с открытым кодом, основанная на теории ARTM. Она имеет расширяемый встроенный набор регуляризаторов и метрик качества, реализует онлайнный и оффлайнный многопоточный пакетный EM-алгоритм, обеспечивающий высокую эффективность обработки больших коллекций на одном компьютере. Библиотека является кроссплатформенной: сборку и исполнение можно производить под Windows 7/8/10, Mac OS и различными дистрибутивами Linux. Поддерживаются программные интерфейсы под Python 2.7.* / 3.*, C++, а также запуск в виде исполняемого бинарного файла. Исходный код BigARTM написан на C++11. Поддерживается несколько популярных форматов текстовых данных. Исчерпывающую информацию по библиотеке можно найти в документации на сайте <http://bigartm.org>.

Ниже представлен минимальный код в Python, выполняющий загрузку и преобразование данных во внутренний формат пакетов документов (*батчей*), создание и обучение модели, вычисление и вывод перплексии.

```
1 # Import all necessary tools and data
2 from sklearn.feature_extraction.text import CountVectorizer
3 from sklearn.datasets import fetch_20newsgroups
4 from numpy import array
5 import artm
6 # Extract data using sklearn and numpy
7 cv = CountVectorizer(max_features=1000, stop_words='english')
8 n_wd = array(cv.fit_transform(fetch_20newsgroups().data).todense()).T
9 vocabulary = cv.get_feature_names()
10 # Create batches and dictionary
11 bv = artm.BatchVectorizer(data_format='bow_n_wd',
12                           n_wd=n_wd,
13                           vocabulary=vocabulary)
14 # Learn simple PLSA model
15 model = artm.ARTM(num_topics=15, dictionary=bv.dictionary)
16 model.scores.add(artm.PerplexityScore(name='perp',
17                                       dictionary=bv.dictionary))
18 model.fit_offline(bv, num_collection_passes=20)
19 # Print perplexity values by iterations
20 print(model.score_tracker['perp'].value)
```

Подготовка данных. Универсальным объектом, принимаемым на вход всеми операциями BigARTM, является векторизатор `artm.BatchVectorizer`. В примере выше (шаги 10–13) он был создан по матрице «мешка слов» `n_wd` и словаря, задающего соответствие между строками матрицы и словами коллекции. В этом случае пакеты создаются в оперативной памяти и полностью удаляются из неё по завершении работы библиотеки. Этот способ хранения данных подходит только для небольших коллекций, целиком помещающихся в памяти. Во всех остальных случаях используются форматы данных, предполагающие чтение исходных документов с диска и запись итоговых пакетов на диск. Наиболее популярен формат текстовых файлов Vowpal Wabbit, в котором каждая строка соответствует одному документу и имеет вид

```
doc_title token_1:value_1 token_2:value_2 ...
```

Данный формат позволяет представлять документы как «мешком слов», так и последовательным текстом, а также записывать в документы токены различных модальностей. Пример создания векторизатора по данным в формате Vowpal Wabbit:

```
1 bv = artm.BatchVectorizer(data_path='docword.vw.txt',
2                             data_format='vowpal_wabbit',
3                             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:

```
1 bv = artm.BatchVectorizer(data_path='my_collection_batches',
2                             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:

```
1 bv = artm.BatchVectorizer(data_path='docword.vw.txt',
2                             data_format='vowpal_wabbit',
3                             target_folder='my_collection_batches')
4 bv.dictionary.save('my_collection_batches/dictionary')
5 # Load dictionary back during next BigARTM launch:
6 dictionary = artm.Dictionary()
7 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:

```
1 dictionary.save_text('my_collection_batches/dictionary.txt')
2 # Change file according to your needs ...
```

```

3 # Then, load it back
4 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 встречаемости 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 (23) с небольшим обобщением:

$$\varphi_{wt} = \operatorname{norm}_{w \in W}(n_{wt} + \tau \beta_w f(\varphi_{wt})).$$

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 (24), с аналогичным обобщением:

$$\theta_{td} = \operatorname{norm}_{t \in T}(n_{td} + \tau \alpha_i \alpha_{td} f(\theta_{td})).$$

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 (25). От пользователя требуется указать коэффициент регуляризации τ и список модальностей, на которые нужно воздействовать.

`TopicSelectionThetaRegularizer` — регуляризатор отбора тем в матрице Θ , реализован по формуле (46). 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 (35). Для корректной работы регуляризатора требуется указать имя модальности времени и расположить токены времени в словаре и в матрице Φ в хронологическом порядке.

`NetPlsaPhiRegularizer` — регуляризатор *NetPLSA* для модальности вершин графа в матрице Φ , определяется по формуле (49). В документах должны быть заранее записаны токены 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 (52) и в качестве параметра требует словарь совстречаемости слов C_{uv} (собрать его можно с помощью встроенного парсера).

`BiternsPhiRegularizer` — регуляризатор битермов, реализован по формулам (50)–(51) 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 (44). 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 (48).

`TopicSegmentationPtdwRegularizer` — регуляризатор *E*-шага для разреживания сегментов в матрицах (p_{tdw}), определяемый по формуле (65).

`SmoothPtdwRegularizer` — регуляризатор *E*-шага для сглаживания матриц (p_{tdw}) по локальному контексту. Приближает тематический профиль каждого вхождения термина 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:

```
1 reg = artm.DecorrelatorPhiRegularizer(name='decor', tau=1e+5)
2 model.regularizer.add(reg)
3 model.scores.SparsityPhiScore(name='sparse')
4
5 model.fit_offline(batch_vectorizer=bv, num_collection_passes=10)
6 print model.score_tracker('sparse').last_value
7
8 # Printing result: 0.15 - too small. Let's increase tau
```

```

9 model.regularizer['decor'].tau = 3e+5
10 model.fit_offline(batch_vectorizer=bv, num_collection_passes=15)

```

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 (66). 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):

```

1 # m = artm.ARTM(...)
2 m.scores.add(artm.PerplexityScore(name='perp',
3                                 class_ids=['@default_class'],
4                                 dictionary=dictionary))

```

Значения перплексии можно вывести следующим образом (вместо `value` можно вывести, например, числитель и знаменатель перплексии по каждой модальности):

```

1 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, содержащий данные о попарной встречаемости слов, то будет вычислена когерентность $coher_t$ по спискам топ-слов в темах, согласно формуле (67).

`TopicKernelScore` — *ядровые характеристики тем*: чистота pur_t , контрастность con_t , размер ядра ker_t оценивающие различность и, косвенно, интерпретируемость каждой темы t , см. стр. 53. Аналогично топ-словам, указание словаря встречаемостей 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):

```
1 model.get_phi(topic_names=model.topic_names[: 10],
2               class_ids=['@default_class'],
3               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:

```
1 # case 2
2 model.get_theta()
3 # case 3
4 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.

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 , $b = 1, \dots, B$, каждый из которых хранится в отдельном файле. Пакеты обрабатываются по очереди. Каждый пакет загружается в память, обновляет матрицу Φ и выгружается. 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.

Функция `ProcessBatches` обрабатывает за один раз множество пакетов $\{D_b\}$, см. Алгоритм 3. 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.

В библиотеке реализовано несколько стратегий агрегирования счётчиков, полученных от разных пакетов, в итоговых счётчиках n_{wt} .

Оффлайнный 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 произ-

Алгоритм 3. Оффлайновый и онлайнвый EM-алгоритм для ARTM

- 1 **функция** $(\tilde{n}_{wt}) := \text{ProcessBatches}$ (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 $p_{tdw} := \text{norm}_{t \in T}(\varphi_{wt}\theta_{td})$ для всех $w \in d, t \in T$;
 - 7 $\theta_{td} := \text{norm}_{t \in T}(\sum_{w \in d} n_{dw}p_{tdw} + \theta_{td}\frac{\partial R}{\partial \theta_{td}})$ 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 $D = \{D_b: b \in B\}$);
 - 11 инициализировать φ_{wt} для всех $w \in W, t \in T$;
 - 12 **повторять**
 - 13 $(n_{wt}) := \sum_{b=1}^B \text{ProcessBatches}(D_b, \Phi)$;
 - 14 $\varphi_{wt} := \text{norm}_{w \in W}(n_{wt} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}})$ для всех $w \in W, t \in T$;
 - 15 **пока** Φ не сойдётся;

 - 16 **функция** FitOnline (коллекция $D = \{D_b: b \in B\}$, параметры $k_{\text{decay}}, k_{\text{apply}}, \eta$);
 - 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** $i := 1, \dots, \lfloor B/\eta \rfloor$
 - 20 $(\tilde{n}_{wt}) := \text{ProcessBatches}(\{D_{\eta(i-1)+1}, \dots, D_{\eta i}\}, \Phi)$;
 - 21 $n_{wt} := k_{\text{decay}}n_{wt} + k_{\text{apply}}\tilde{n}_{wt}$ для всех $w \in W, t \in T$;
 - 22 $\varphi_{wt} := \text{norm}_{w \in W}(n_{wt} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}})$ для всех $w \in W, t \in T$;
-

водится в конце каждого прохода коллекции в однопоточном режиме. Оффлайновый режим ориентирован на обработку относительно небольших коллекций.

Онлайнный алгоритм FitOnline был предложен для модели LDA в [48], позже для модели PLSA в [20]. Он реализован в библиотеках машинного обучения Vowpal Wabbit, Gensim 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. Онлайнный алгоритм делает один проход по коллекции, обновляя матрицу Φ после каждых η пакетов: на шаге 21 счётчики \tilde{n}_{wt} , накопленные по η последним пакетам, суммируются со счётчиками n_{wt} , накопленными по всем пакетам, с весами k_{decay} и k_{apply} . На шаге 22 матрица Φ пересчитывается по обновлённым счётчикам с учётом регуляризаторов. Онлайнный алгоритм ориентирован на обработку больших коллекций или потоков данных.

Весовые коэффициенты k_{decay} и k_{apply} позволяют управлять темпом забывания предыдущих пакетов. Имеется возможность задать эти коэффициенты по формулам,

рекомендованным в [48]: $k_{\text{decay}} = 1 - \rho_i$, $k_{\text{apply}} = \rho_i$, 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.

Контроль условий сходимости EM-алгоритма возлагается на пользователя. Проще всего задавать число итераций по коллекции и по каждому документу.

В онлайн-алгоритме разбиение коллекции на пакеты и порядок обработки пакетов могут влиять на результат, в отличие от оффлайн-алгоритма. Чтобы уменьшить это влияние, коллекцию разбивают на пакеты случайным образом.

В онлайн-алгоритме, как и в оффлайн-алгоритме, внешний цикл по пакетам выполняется параллельно в несколько потоков. При этом обработка следующей порции пакетов не начинается, пока не завершено обновление матрицы Φ на шагах 21–22. В результате вычислительные ресурсы используются неэффективно. Эта проблема решается *детерминированным асинхронным онлайн-алгоритмом*, в котором применяются обновления с запаздыванием [44].

В BigARTM оффлайн-алгоритм запускается функцией `ARTM.fit_offline`, онлайн-алгоритм — функцией `ARTM.fit_online`. Для включения детерминированного асинхронного алгоритма последней надо передать параметр `async=True`.

Построенную тематическую модель можно использовать для тематизации отдельных документов при фиксированной матрице Φ . Эта возможность реализуется функцией `ARTM.transform`, которая пропускает документы через `ProcessBatches`.

Иерархическая тематическая модель реализована в BigARTM отдельным классом `hARTM`. Уровни иерархии создаются функцией `add_level`, для каждого уровня задается число тем `num_topics`. Функция `add_level` возвращает объект `ARTM`, с которым можно работать как с обычной «плоской» моделью, в частности, добавлять к ней регуляризаторы. Пример создания двухуровневой модели:

```
1 hier = artm.hARTM()
2 level0 = hier.add_level(num_topics=5)
3 level1 = hier.add_level(num_topics=25, parent_level_weight=1)
```

Функция `hier.fit_offline` обучает тематическую иерархию по заданной коллекции, используя регуляризатор (47) для связывания каждого уровня иерархии с родительским. Матрицы Φ и Θ всех уровней могут быть получены с помощью функций `hier.get_phi` и `hier.get_theta`.

16 Разведочный информационный поиск

Важным приложением тематического моделирования является *информационный поиск* (information retrieval) [154, 14]. 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) [74, 148]. 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 [9]. 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 [121, 21, 98, 28, 13, 140].

В литературе по разведочному поиску тематическое моделирование стали использовать относительно недавно [112, 46, 109, 129], а многие обзоры о нём вообще не упоминают [42, 106, 119, 55, 75, 52]. В недавней статье [129] важными преимуществами тематических моделей называются гибкость, возможности визуализации и навигации. В то же время, в качестве недостатков отмечают 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; онлайн-алгоритмы способны обрабатывать большие коллекции и потоки документов за линейное время [81, 20, 131]. С другой стороны, в работах по тематическому моделированию разведочный поиск часто называют одним из важнейших приложений, а оценки качества поиска используют для валидации моделей [154, 14]. 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 на-

вигации 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 мультязычной 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 темпоральной, 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 ассессоров или логам пользователей, чтобы постоянно улучшать качество поиска. Наконец, реализация должна быть онлайн-овой, 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 [12] и улучшена в [152]. Для измерения качества разведочного тематического поиска использовались критерии точности и полноты на основе оценок ассессоров. Для оценивания была составлена выборка запросов — заданий разведочного поиска. Каждый запрос представлял собой текст объёмом около одной страницы формата А4, описывающий тематику поиска. Каждое задание сначала выполнялось независимо несколькими ассессорами, затем системой тематического поиска, затем её результат снова оценивался ассессорами. 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 ассессоры, сокращая среднее время поиска с получаса до секунды. Комбинирование регуляризаторов декоррелирования, разреживания и сглаживания вместе с модальностями n -грамм, авторов и категорий значительно 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 обзора⁶, 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

⁶<http://textvis.lnu.se> — интерактивный обзор средств визуализации текстов.

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 масштабирование и фильтрация, детали по требованию»⁷ [117].

Графическое отображение результатов тематического моделирования и разведочного поиска согласуется с концепцией *дальнего чтения* (*distant reading*), предложенной социологом литературы Франко Моретти [85]. Он противопоставляет этот способ изучения текстов нашему обычному чтению (*close reading*). Невозможно прочитать миллионы книг или статей, но вполне возможно применить статистические методы и графическую визуализацию, чтобы понять в общих чертах, о чём вся эта литература, 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»⁸.

Для библиотеки **BigARTM** в настоящее время развивается инструмент визуализации с веб-интерфейсом **VisARTM**⁹, 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.

17 Заключение

Тематическое моделирование является одним из инструментов статистического анализа текстов. За полтора десятилетия интенсивных исследований созданы сотни тематических моделей. Основные их виды включены 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 определяется простой вероятностной порождающей моделью, описывающей структуру пространства параметров, и «мешком регуляризаторов», задающих дополнительные требования. Регуляризаторы аддитивны, взаимозаменяемы и легко отторгаются от моделей. Их можно комбинировать практически в любых сочетаниях, что и приводит к модульной технологии моделирования. Создание модели с требуемыми свойствами под конкретное приложение не требует ни трудоёмких математических выкладок, ни создания алгоритмически сложного кода.

⁷Visual Information Seeking Mantra: «Overview first, zoom and filter, details on demand» [117].

⁸«*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.» [85].

⁹Федоряка Д. С. Технология интерактивной визуализации тематических моделей. Бакалаврская диссертация, МФТИ, 2017 (www.MachineLearning.ru/wiki/images/d/d8/Fedoriaka17bsc.pdf).

Главное отличие байесовского обучения в том, что там дополнительные требования встраиваются в порождающую модель, 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.

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 подхода.
- «Изучение тематического моделирования надо начинать с графической нотации (plate notation), распределения Дирихле и байесовского вывода». Не обязательно. Проще разобраться, что такое матричное разложение и регуляризация.
- «Тематические модели предсказывают слова 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, бигермов, предложений, сегментации, регуляризаторы E -шага позволяют учитывать порядок слов в документах.
- «Тематические модели проигрывают конкуренцию с word2vec». Модели бигермов и сети слов строятся по данным о встречаемости слов, как и word2vec. Они вполне конкурентоспособны, а тематические векторные представления можно сделать интерпретируемыми, разреженными, иерархическими.
- «Тематическая модель LDA переобучается гораздо меньше, чем PLSA». Нет, они примерно одинаковы, особенно на больших коллекциях. LDA лучше описывает вероятности редких слов, которые наименее важны для образования тем, но формально это ухудшает правдоподобие модели.
- «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 ускорить сходимость регуляризованного EM-алгоритма? Как подбирать коэффициенты регуляризации адаптивно и полностью автоматически в ходе итераций? 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 именованя 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»¹⁰, который автор читает на факультете ВМК Московского Государственного Университета им. М. В. Ломоносова. Обновляемая электронная версия доступна на сайте MachineLearning.ru¹¹.

Работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований (проект № 17-07-01536) 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://jmla.org>)*. — 2015. — Т. 1, № 11. — С. 1584–1618.
- [3] Воронцов К. В. Аддитивная регуляризация тематических моделей коллекций текстовых документов // *Доклады РАН*. — 2014. — Т. 456, № 3. — С. 268–271.
- [4] Воронцов К. В., Потапенко А. А. Регуляризация, робастность и разреженность вероятностных тематических моделей // *Компьютерные исследования и моделирование*. — 2012. — Т. 4, № 4. — С. 693–706.
- [5] 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.
- [6] 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.
- [7] Дударенко М. А. Регуляризация многоязычных тематических моделей // *Вычислительные методы и программирование*. — 2015. — Т. 16. — С. 26–38.
- [8] Лукашевич Н. В. Тезаурусы в задачах информационного поиска. — Издательство МГУ имени М. В. Ломоносова, 2011.
- [9] Маннинг К. Д., Рагхаван П., Шютце Х. Введение в информационный поиск. — Вильямс, 2011.

¹⁰<http://www.MachineLearning.ru/wiki?title=BTM>.

¹¹<http://www.MachineLearning.ru/wiki/images/d/d5/Voron17survey-artm.pdf>.

- [10] *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.
- [11] *Тихонов А. Н., Арсенин В. Я.* Методы решения некорректных задач. — М.: Наука, 1986.
- [12] *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.
- [13] *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.
- [14] *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.
- [15] *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.
- [16] *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.
- [17] *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.
- [18] *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.
- [19] *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.
- [20] *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.
- [21] *Blei D., Lafferty J.* A correlated topic model of Science // *Annals of Applied Statistics.* — 2007. — Vol. 1. — Pp. 17–35.
- [22] *Blei D. M.* Probabilistic topic models // *Communications of the ACM.* — 2012. — Vol. 55, no. 4. — Pp. 77–84.
- [23] *Blei D. M., Griffiths T., Jordan M., Tenenbaum J.* Hierarchical topic models and the nested chinese restaurant process // *NIPS.* — 2003.
- [24] *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.
- [25] *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.
- [26] *Blei D. M., Ng A. Y., Jordan M. I.* Latent Dirichlet allocation // *Journal of Machine Learning Research.* — 2003. — Vol. 3. — Pp. 993–1022.

- [27] *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.
- [28] *Boelli 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.
- [29] *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.
- [30] *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.
- [31] *Chen B.* Word topic models for spoken document retrieval and transcription. — 2009. — Vol. 8, no. 1. — Pp. 2:1–2:27.
- [32] *Chien J.-T., Chang Y.-L.* Bayesian sparse topic model // *Journal of Signal Processing Systems*. — 2013. — Vol. 74. — Pp. 375–389.
- [33] *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.
- [34] *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.
- [35] *Dai A. M., Olah C., Le Q. V.* Document embedding with paragraph vectors // NIPS Deep Learning Workshop.
- [36] *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.
- [37] *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.
- [38] *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.
- [39] *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.
- [40] *Eisenstein J., Ahmed A., Xing E. P.* Sparse additive generative models of text // ICML'11. — 2011. — Pp. 1041–1048.
- [41] *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.
- [42] *Feldman S. E.* The answer machine // Synthesis Lectures on Information Concepts, Retrieval, and Services. — Morgan & Claypool Publishers, 2012. — Vol. 4. — Pp. 1–137.
- [43] *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.
- [44] *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.

- [45] *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 information retrieval. — 2003. — Pp. 433–434.
- [46] *Grant C. E., George C. P., Kanjilal V., Nirakhiwale S., Wilson J. N., Wang D. Z.* A topic-based search, visualization, and exploration system // FLAIRS Conference. — AAAI Press, 2015. — Pp. 43–48.
- [47] *Harris Z.* Distributional structure // *Word*. — 1954. — Vol. 10, no. 23. — Pp. 146–162.
- [48] *Hoffman M. D., Blei D. M., Bach F. R.* Online learning for latent Dirichlet allocation // NIPS. — Curran Associates, Inc., 2010. — Pp. 856–864.
- [49] *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.
- [50] *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.
- [51] *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.
- [52] *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.
- [53] *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.
- [54] *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.
- [55] *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.
- [56] *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.
- [57] *Kochedykov D., Apishev M., Golitsyn L., Vorontsov K.* 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. — P. 182–193.
- [58] *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.
- [59] *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.
- [60] *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.
- [61] *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.

- [62] *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.
- [63] *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.
- [64] *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.
- [65] *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.
- [66] *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.
- [67] *Li W., McCallum A.* Pachinko allocation: Dag-structured mixture models of topic correlations // ICML. — 2006.
- [68] *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.
- [69] *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.
- [70] *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.
- [71] *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.
- [72] *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.
- [73] *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.
- [74] *Marchionini G.* Exploratory search: From finding to understanding // *Commun. ACM.* — 2006. — Vol. 49, no. 4. — Pp. 41–46.
- [75] *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.
- [76] *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.
- [77] *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.
- [78] *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.

- [79] *Mikolov T., Chen K., Corrado G., Dean J.* Efficient estimation of word representations in vector space // *CoRR*. — 2013. — Vol. abs/1301.3781.
- [80] *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.
- [81] *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.
- [82] *Mimno D., Li W., McCallum A.* Mixtures of hierarchical topics with pachinko allocation // ICML. — 2007.
- [83] *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.
- [84] *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.
- [85] *Moretti F.* Graphs, maps, trees : abstract models for literary history. — London; New York: Verso, 2007.
- [86] *Nadeau D., Sekine S.* A survey of named entity recognition and classification // *Linguisticae Investigationes*. — 2007. — Vol. 30, no. 1. — Pp. 3–26.
- [87] *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.
- [88] *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.
- [89] *Newman D., Karimi S., Cavedon L.* External evaluation of topic models // Australasian Document Computing Symposium. — December 2009. — Pp. 11–18.
- [90] *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.
- [91] *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.
- [92] *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.
- [93] *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.
- [94] *Nikolenko S. I., Koltcov S., Koltsova O.* Topic modelling for qualitative studies // *Journal of Information Science*. — 2017. — Vol. 43, no. 1. — Pp. 88–102.
- [95] *Pagliardini M., Gupta P., Jaggi M.* Unsupervised learning of sentence embeddings using compositional n -gram features // *CoRR*. — 2017. — Vol. abs/1703.02507.

- [96] *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.
- [97] *Paul M. J., Dredze M.* Discovering health topics in social media using topic models // *PLoS ONE.* — 2014. — Vol. 9, no. 8.
- [98] *Paul M. J., Girju R.* Topic modeling of research fields: An interdisciplinary perspective // *RANLP. — RANLP 2009 Organising Committee / ACL, 2009.* — Pp. 337–342.
- [99] *Pennington J., Socher R., Manning C. D.* Glove: Global vectors for word representation // *Empirical Methods in Natural Language Processing (EMNLP).* — 2014. — Pp. 1532–1543.
- [100] *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.
- [101] *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.
- [102] *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.
- [103] *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.
- [104] *Pritchard J. K., Stephens M., Donnelly P.* Inference of population structure using multilocus genotype data // *Genetics.* — 2000. — Vol. 155. — Pp. 945–959.
- [105] *Pujara J., Skomoroch P.* Large-scale hierarchical topic models // *NIPS Workshop on Big Learning.* — 2012.
- [106] *Rahman M.* Search engines going beyond keyword search: A survey // *International Journal of Computer Applications.* — August 2013. — Vol. 75, no. 17. — Pp. 1–8.
- [107] *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.
- [108] *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.
- [109] *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.
- [110] *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.
- [111] *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.
- [112] *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, TPD 2013, Valletta, Malta, September 22-26, 2013. Proceedings / Ed. by T. Aalberg, C. Papatheodorou, M. Dobрева, G. Tsakonas, C. J. Farrugia.* — Berlin, Heidelberg: Springer Berlin Heidelberg, 2013. — Pp. 370–373.

- [113] *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.
- [114] *Sharma A., Pawar D. M.* Survey paper on topic modeling techniques to gain usefull forecasting 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.
- [115] *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.
- [116] *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.
- [117] *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.
- [118] *Si X., Sun M.* Tag-LDA for scalable real-time tag recommendation // *Journal of Information & Computational Science*. — 2009. — Vol. 6. — Pp. 23–31.
- [119] *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.
- [120] *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.
- [121] *Steyvers M., Griffiths T.* Finding scientific topics // *Proceedings of the National Academy of Sciences*. — 2004. — Vol. 101, no. Suppl. 1. — Pp. 5228–5235.
- [122] *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.
- [123] *Tan Y., Ou Z.* Topic-weak-correlated latent Dirichlet allocation // *7th International Symposium Chinese Spoken Language Processing (ISCSLP)*. — 2010. — Pp. 224–228.
- [124] *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.
- [125] *Teh Y. W., Newman D., Welling M.* A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation // *NIPS*. — 2006. — Pp. 1353–1360.
- [126] *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.
- [127] *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.
- [128] *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.
- [129] *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, CCAHI2015, International Joint Conference on Artificial Intelligence, IJCAI, Buenos Aires, Argentina, July 2015*. — Aachen, Germany, Germany: CEUR-WS.org, 2015.

- [130] Vorontsov K., Frei O., Apishev M., Romov P., Suvorova M. 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.
- [131] 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.
- [132] Vorontsov K. V., Potapenko A. A. Additive regularization of topic models // *Machine Learning, Special Issue on Data Analysis and Intelligent Optimization*. — 2014.
- [133] 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.
- [134] 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.
- [135] 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.
- [136] 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.
- [137] 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.
- [138] 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.
- [139] Wang C., Blei D. M. Decoupling sparsity and smoothness in the discrete hierarchical Dirichlet process // NIPS. — Curran Associates, Inc., 2009. — Pp. 1982–1989.
- [140] 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.
- [141] 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.
- [142] 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.
- [143] Wang C., Liu X., Song Y., Han J. Scalable and robust construction of topical hierarchies // *CoRR*. — 2014. — Vol. abs/1403.3460.
- [144] 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.

- [145] 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.
- [146] 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.
- [147] 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.
- [148] 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.
- [149] 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.
- [150] 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.
- [151] 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.
- [152] 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. — Springer, Cham, 2017. — P. 181–193.
- [153] 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.
- [154] 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.
- [155] 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.
- [156] Yin H., Cui B., Sun Y., Hu Z., Chen L. LCARS: A spatial item recommender system // *ACM Transaction on Information Systems*. — 2014.
- [157] 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.
- [158] 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.
- [159] 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.
- [160] 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.

- [161] *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.
- [162] *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.
- [163] *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.