Тематическое моделирование в BigARTM: теория, алгоритмы, приложения

К. В. Воронцов, А. И. Фрей, М. А. Апишев, А. А. Потапенко, ...

14 июня 2015 г.

Содержание

1 Введение

Тематическое моделирование — это одно из современных направлений статистического анализа текстов, активно развивающееся с конца 90-х годов. *Вероятностная тематическая модель* (probabilistic topic model) коллекции текстовых документов описывает каждую тему дискретным распределением на множестве терминов, каждый документ — дискретным распределением на множестве тем.

Одним из основных приложений тематического моделирования является *информационный поиск* (information retrieval) [74, 10]. Современные поисковые системы предназначены для поиска по коротким текстовым запросам. Они основаны на инвертированных индексах, в которых для каждого слова хранится список содержащих его документов [6]. Поисковая система ищет документы, содержащие все слова запроса, поэтому по длинному запросу, скорее всего, ничего не будет найдено. Тематический или *разведочный поиск* (exploratory search, information seeking) — это разновидность информационного поиска. Он более подходит не для ответов на конкретные вопросы, а для расширения профессиональных знаний. Если пользователь плохо ориентируется в терминологии или слабо представляет себе структуру предметной области, то его информационной потребностью является скорее получение «дорожной карты» предметной области, систематизация и визуализация релевантной информации по заданной теме. Тема запроса не формулируется словами, а задаётся текстовым фрагментом произвольный длины. Поисковая система 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. Новые технологии информационного поиска на основе тематического моделирования в настоящее время активно разрабатываются [56, 13, 45, 18, 9, 70].

Тематические модели применяются также для выявления трендов в новостных потоках или научных публикациях [78, 59], для многоязычного информационного поиска [67, 66], для анализа данных социальных сетей [79, 61, 47], для классификации и категоризации документов [51, 80], для тематической сегментации текстов [71], для анализ изображений и видеопотоков [28, 36, 25, 60], для тегирования веб-страниц [32], для обнаружения текстового спама [7], для рекомендательных систем [73, 70, 35, 76, 75], для анализа нуклеотидных [33] и аминокислотных последовательностей [53, 31], в задачах популяционной генетики [49].

Построение тематической модели по коллекции документов сводится к оптимизационной задаче стохастического матричного разложения. В общем случае она имеет бесконечное множество решений, то есть является некорректно поставленной. Согласно теории регуляризации А. Н. Тихонова [8], если задача недоопределена, то её решение можно сделать устойчивым, добавив к основному критерию дополнительный критерий — регуляризатор, учитывающий специфику предметной области.

Аддитивная регуляризация тематических моделей (additive regularization for topic modeling, ARTM) — это многокритериальный подход, в котором к основному критерию добавляется взвешенная сумма регуляризаторов [1, 63, 4, 62]. Тематическое моделирование отличается от других областей машинного обучения большим разнообразием регуляризаторов. Многие из них исходно предлагались как самостоятельные модели. АРТМ позволяет комбинировать тематические модели, суммируя регуляризаторы. Благодаря свойству аддитивности, оптимизация любых моделей и

их комбинаций производится одним и тем же итерационным процессом, называемым EM-алгоритмом. Для добавления регуляризатора в модель достаточно знать его частные производные по параметрам модели. EM-алгоритм хорошо масштабируется, поскольку каждая его итерация — это один линейный проход по коллекции, а число итераций, требуемых для сходимости процесса, как правило, невелико.

Подчеркнём, что APTM — это не ещё одна тематическая модель, а общий подход к построению и комбинированию многих тематических моделей.

Онлайновый EM-алгоритм реализует очень эффективную схему вычислений, при которой большие коллекции документов могут обрабатываться вообще за одну итерацию. Это происходит благодаря тому, что в матричном разложении $\Phi\Theta$ матрица Φ зависит от всей коллекции, тогда как в матрице Θ каждый столбец зависит от одного отдельного документа. Если коллекция настолько избыточна, что её тематика уже неплохо определяется по небольшой доле документов, то матрица Φ успевает сойтись задолго до того, как заканчивается первая итерация.

Параллельный онлайновый EM-алгоритм также эксплуатирует идею избыточности коллекции. Коллекция разделяется на пакеты, которые обрабатываются параллельно на разных ядрах или на разных узлах вычислительной сети. Для каждого пакета вычисляются тематические модели составляющих его документов, при этом обновления общей матрицы Φ производятся периодически и довольно редко, после чего обновлённая матрица Φ рассылается всем узлам. Это позволяет обрабатывать сколь угодно большие коллекции, которые не помещаются целиком ни в оперативную память, ни даже на диск одного компьютера.

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 для предсказания пропущенных метаданных. Примерами метаданных являются: авторы [50], моменты времени [59, 78, 60], классы, жанры или категории [51, 80], цитируемые или цитирующие документы [23] или авторы [30], пользователи электронных библиотек, социальных сетей или рекомендательных систем [35, 54, 70, 75, 76], графические элементы изображений [16, 28, 36], рекламные объявления на веб-страницах [46]. Каждый тип метаданных образует отдельную модальность со своим словарём. Слова, составляющие основной текст документов, образуют лишь одну из модальностей. Словосочетания (n-граммы, коллокации), образующие альтернативное представление текста, могут вводиться в тематическую модель отдельной модальностью. Для анализа коротких текстов используют битермы — пары слов, употреблённых в одном сообщении, но не обязательно стоящих рядом [72]. Для анализа коротких текстов с опечатками используют буквенные n-граммы, что позволяет улучшать качество информационного поиска [29]. Тэги [32] и именованные сущности (named entities) [41], хотя и обозначаются в текстах словами, но имеют более чёткую семантику, поэтому выделяются в отдельную модальность. В коллекциях с параллельными текстами на нескольких языках модальностями являются языки [5, 21, 37, 44, 67]. Мультимодальные аддитивно регуляризованные тематические модели обобщают эти и другие случаи, а также их возможные комбинации, позволяя обрабатывать метаданные с произвольным числом модальностей.

BigARTM — это библиотека тематического моделирования с открытым кодом, доступная по адресу `http://bigartm.org`. Она реализует все перечисленные выше

возможности, включает в себя набор регуляризаторов и набор метрик качества тематических моделей. Решение разнообразных прикладных задач с большим объёмом сложно структурированных данных становится возможным в BigARTM благодаря свойству аддитивности регуляризаторов. Пользователь выбирает набор регуляризаторов, соответствующих целям моделирования, затем экспериментально подбирает *стратегию регуляризации*, то есть определяет, по какому закону будут меняться весовые коэффициенты регуляризаторов.

Согласно теории регуляризации А. Н. Тихонова, коэффициенты регуляризации необходимо устремлять к нулю в ходе итераций для получения устойчивого решения исходной нерегуляризованной задачи. На практике одни регуляризаторы могут выполнять подготовительную работу для других, поэтому важна последовательность и постепенность включения и отключения регуляризаторов. Автоматический выбор стратегии регуляризации пока остаётся открытой теоретической проблемой. В текущей версии BigARTM используются относительные коэффициенты регуляризации, облегчающие перенос отработанных стратегий регуляризации с одной задачи на другую независимо от их размерных характеристик.

В данном документе описывается теория, лежащая в основе BigARTM, набор регуляризаторов, набор метрик качества, некоторые эксперименты и рекомендации по выбору стратегий регуляризации.

2 Вероятностное тематическое моделирование

Пусть D — множество (коллекция) текстовых документов, W — множество (словарь) всех употребляемых в них терминов. Терминами могут быть как отдельные слова, так и словосочетания. Каждый документ $d \in D$ представляет собой последовательность n_d терминов $w_1, \ldots, w_{n_d}$ из словаря W.

Вероятностное пространство. Предполагается, что существует конечное множество тем T, и каждое вхождение термина w в документ d связано с некоторой темой $t \in T$. Коллекция документов рассматривается как случайная и независимая выборка троек (w_i, d_i, t_i), $i = 1, \ldots, n$ из дискретного распределения $p(w, d, t)$ на конечном вероятностном пространстве $W \times D \times T$. Термины w и документы d являются наблюдаемыми переменными, тема $t \in T$ является *латентной* (скрытой) переменной.

Гипотеза мешка слов. Предполагается, что порядок терминов в документах не важен для выявления тематики, то есть тематику документа можно узнать даже после произвольной перестановки терминов, хотя для человека такой текст теряет 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 подмножества $d \subset W$, каждому элементу которого, $w \in d$, поставлено в соответствие число n_{dw} вхождений токена w в документ d.

Гипотеза условной независимости. Предполагается, что появление слов в документе d по теме t зависит от темы, но не зависит от документа d, и описывается общим для всех документов распределением $p(w \mid t)$. Это предположение, называемое

Алгоритм 2.1. Вероятностная модель порождения коллекции документов.

Вход: распределения $p(w \mid t)$, $p(t \mid d)$;
Выход: выборка пар (d_i, w_i), $i = 1, \ldots, n$;
для всех $d \in D$
  задать длину n_d документа d;
  для всех $i = 1, \ldots, n_d$
    $d_i := d$;
    выбрать случайную тему t_i из распределения $p(t \mid d_i)$;
    выбрать случайный термин w_i из распределения $p(w \mid t_i)$;

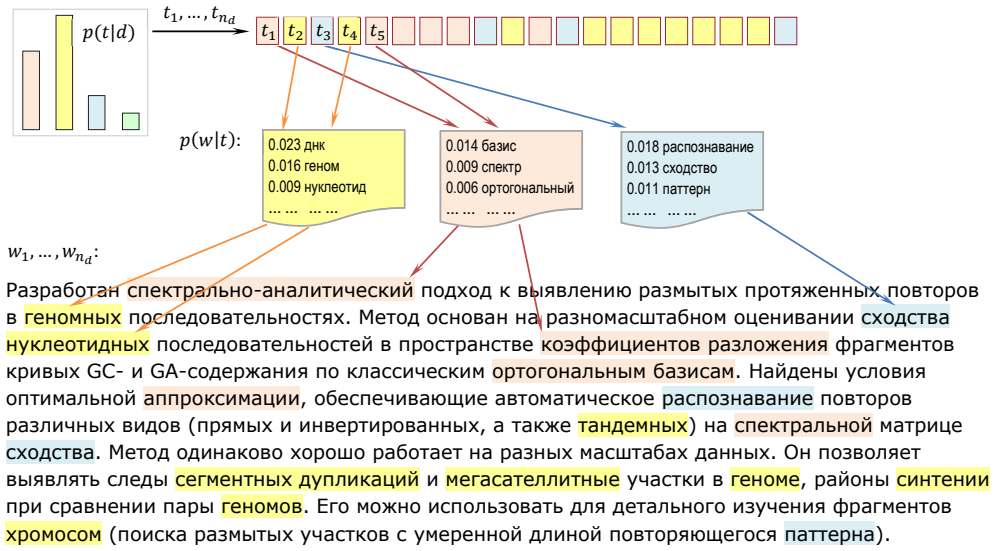


Рис. 1. Процесс порождения текстового документа вероятностной тематической моделью (2.2).

гипотезой условной независимости, допускает три эквивалентных представления:

$$\begin{aligned} &p(w \mid d, t) = p(w \mid t); \\ &p(d \mid w, t) = p(d \mid t); \\ &p(d, w \mid t) = p(d \mid t) p(w \mid t). \end{aligned} \tag{2.1}$$

Вероятностная порождающая модель выражает вероятности $p(w \mid d)$ появления терминов w в документах d через распределения $p(w \mid t)$ и $p(t \mid d)$. Согласно формуле полной вероятности и гипотезе условной независимости,

$$p(w \mid d) = \sum_{t \in T} p(w \mid t)\, p(t \mid d). \tag{2.2}$$

Вероятностная модель (2.2) описывает процесс порождения коллекции D по известным распределениям $p(w \mid t)$ и $p(t \mid d)$, см. Алгоритм 2.1 и рис. 1.

Построение тематической модели — это обратная задача: по известной коллекции D требуется оценить параметры модели $\varphi_{wt} = p(w \mid t)$ и $\theta_{td} = p(t \mid d)$.

Обычно число тем $|T|$ много меньше $|D|$ и $|W|$, и задача сводится к поиску приближённого представления заданной матрицы частот $F = (f_{wd})_{W \times D}$, $f_{wd} = \frac{n_{dw}}{n_d}$

в виде произведения $F \approx \Phi\Theta$ двух неизвестных матриц меньшего размера — *матрицы терминов тем* $\Phi = (\varphi_{wt})_{W\times T}$ и *матрицы тем документов* $\Theta = (\theta_{td})_{T\times D}$. Все три матрицы F, Φ, Θ являются *стохастическими*, то есть имеют неотрицательные нормированные столбцы f_d, φ_t, θ_d, представляющие дискретные распределения.

Частотные оценки условных вероятностей. Вероятности, связанные с наблюдаемыми переменными d, w, можно оценивать по выборке как частоты. Частотные оценки являются несмещёнными оценками максимального правдоподобия (здесь и далее выборочные оценки вероятностей p будем обозначать через $\hat{p}$):

$$\hat{p}(d,w) = \frac{n_{dw}}{n}, \qquad \hat{p}(d) = \frac{n_d}{n}, \qquad \hat{p}(w) = \frac{n_w}{n}, \qquad \hat{p}(w \mid d) = \frac{n_{dw}}{n_d}, \tag{2.3}$$

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

$n_d = \sum\limits_{w\in W} n_{dw}$ — длина документа d в терминах;

$n_w = \sum\limits_{d\in D} n_{dw}$ — число вхождений термина w во все документы коллекции;

$n = \sum\limits_{d\in D}\sum\limits_{w\in d} n_{dw}$ — длина коллекции в терминах.

Вероятности, связанные со скрытой переменной t, также можно оценивать как частоты, если рассматривать коллекцию документов как выборку троек (d, w, t):

$$\hat{p}(t) = \frac{n_t}{n}, \qquad \hat{p}(w \mid t) = \frac{n_{wt}}{n_t}, \qquad \hat{p}(t \mid d) = \frac{n_{dt}}{n_d}, \qquad \hat{p}(t \mid d, w) = \frac{n_{tdw}}{n_{dw}}, \tag{2.4}$$

n_{tdw} — число троек, в которых термин w документа d связан с темой t;

$n_{dt} = \sum\limits_{w\in W} n_{tdw}$ — число троек, в которых термин документа d связан с темой t;

$n_{wt} = \sum\limits_{d\in D} n_{tdw}$ — число троек, в которых термин w связан с темой t;

$n_t = \sum\limits_{d\in D}\sum\limits_{w\in d} n_{tdw}$ — число троек, связанных с темой t.

Эти оценки не могут быть вычислены непосредственно по исходным данным, так как темы t_i неизвестны. Однако можно заметить, что все оценки в (2.4) выражаются через $n_{tdw} = p(t \mid d, w) n_{dw}$. Зная условные распределения $p(t \mid d, w)$, можно оценить искомые параметры тематической модели $\varphi_{wt} = \hat{p}(w \mid t)$ и $\theta_{td} = \hat{p}(t \mid d)$.

В пределе $n \to \infty$ частотные оценки $\hat{p}(\cdot)$, определяемые формулами (2.3)–(2.4), стремятся к соответствующим вероятностям $p(\cdot)$, согласно закону больших чисел.

При первом знакомстве с вероятностным тематическим моделированием могут возникать трудности с пониманием статистической модели текста и вполне оправданные сомнения в применимости самого понятия «вероятности» к текстам естественного языка. Частотная интерпретация даёт простое и конструктивное понимание всех условных вероятностей, используемых в вероятностных тематических моделях, а также естественных ограничений, присущих этим моделям.

§2.1 Вероятностный латентный семантический анализ

Принцип максимума правдоподобия. Для оценивания параметров Φ, Θ тематической модели по коллекции документов D будем максимизировать правдоподобие

(плотность распределения) выборки:

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

Прологарифмируем правдоподобие, чтобы превратить произведения в суммы, и отбросим константные слагаемые, не зависящие от параметров модели. Получим задачу максимизации логарифма правдоподобия (log-likelihood) при ограничениях неотрицательности и нормированности столбцов матриц Φ и Θ:

$$L(\Phi,\Theta)=\sum_{d\in D}\sum_{w\in d}n_{dw}\ln\sum_{t\in T}\varphi_{wt}\theta_{td}\;\to\;\max_{\Phi,\Theta};\tag{2.5}$$

$$\sum_{w\in W}\varphi_{wt}=1;\qquad\varphi_{wt}\geqslant 0;\tag{2.6}$$

$$\sum_{t\in T}\theta_{td}=1;\qquad\theta_{td}\geqslant 0.\tag{2.7}$$

Данная постановка задачи является основой *вероятностного латентного семантического анализа* (probabilistic latent semantic analysis, PLSA) [27]. Для её решения используется EM-алгоритм, который мы чуть позже выведем строго и в более общей постановке. Сейчас мы придём к этому алгоритму элементарным путём, который даёт его интуитивное понимание.

EM-алгоритм. Искомые параметры модели выражаются через частотные оценки условных вероятностей: $\varphi_{wt}=\frac{n_{wt}}{n_t}$ и $\theta_{td}=\frac{n_{dt}}{n_d}$. Они связаны со скрытыми темами и потому не могут быть вычислены по наблюдаемой коллекции. Тем не менее, их можно оценить, зная $n_{tdw}=n_{dw}p(t\,|\,d,w)$. Чтобы выразить условные вероятности $p(t\,|\,d,w)$ через параметры модели, воспользуемся формулой Байеса:

$$p(t\,|\,d,w)=\frac{p(t,w\,|\,d)}{p(w\,|\,d)}=\frac{p(w\,|\,t)p(t\,|\,d)}{p(w\,|\,d)}=\frac{\varphi_{wt}\theta_{td}}{\sum\limits_{s\in T}\varphi_{ws}\theta_{sd}}.$$

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

$$p_{tdw}=\frac{\varphi_{wt}\theta_{td}}{\sum\limits_{s\in T}\varphi_{ws}\theta_{sd}};\tag{2.8}$$

$$\varphi_{wt}=\frac{n_{wt}}{\sum\limits_{v\in W}n_{vt}};\qquad n_{wt}=\sum_{d\in D}n_{dw}p_{tdw};\tag{2.9}$$

$$\theta_{td}=\frac{n_{td}}{\sum\limits_{s\in T}n_{ds}};\qquad n_{td}=\sum_{w\in d}n_{dw}p_{tdw}.\tag{2.10}$$

Для решения данной системы нелинейных уравнений подходит метод простых итераций: сначала выбираются начальные приближения параметров φ_{wt} и θ_{td}, затем вычисления по формулам (2.8)–(2.10) продолжаются в цикле до сходимости.

Алгоритм 2.2. PLSA-EM: рациональный EM-алгоритм для модели PLSA.

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

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

Применение принципа максимума правдоподобия и решение оптимизационной задачи с помощью EM-алгоритма [22] приводит в PLSA ровно к тому же итерационному процессу [27]. В терминах EM-алгоритма вычисление условных вероятностей по формуле (2.8) называется E-шагом (expectation), а вычисление оценок максимального правдоподобия по формулам (2.9)–(2.10) — M-шагом (maximization).

Рациональный EM-алгоритм. Вычисление переменных n_{wt}, n_{dt}, n_t на M-шаге требует однократного прохода всей коллекции в цикле по всем документам $d \in D$ и всем терминам $w \in d$. Внутри этого цикла переменные p_{tdw} можно вычислять только в тот момент, когда они нужны. От этого результат алгоритма не изменяется, E-шаг встраивается внутрь M-шага без дополнительных вычислительных затрат, отпадает необходимость хранения трёхмерной матрицы p_{tdw}. Этот вариант реализации EM-алгоритма будем называть *рациональным*; он показан в Алгоритме 2.2.

§2.2 Аддитивная регуляризация

Задача стохастического матричного разложения является некорректно поставленной, так как множество её решений в общем случае бесконечно. Если $F = \Phi\Theta$ — решение, то $F = (\Phi S)(S^{-1}\Theta)$ также является решением для всех невырожденных S, при которых матрицы $\Phi' = \Phi S$ и $\Theta' = S^{-1}\Theta$ являются стохастическими.

Существует общий подход к решению некорректно поставленных обратных задач, называемый *регуляризацией* [8]. Когда оптимизационная задача недоопределена, к основному критерию добавляют дополнительный критерий — регуляризатор, по возможности учитывающий специфические особенности данной задачи и знания предметной области.

Аддитивная регуляризация тематических моделей (APTM) [1, 63, 62] основана на введении дополнительных критериев-*регуляризаторов* $R_i(\Phi, \Theta)$, $i = 1, \ldots, r$,

и максимизации их линейной комбинации с логарифмом правдоподобия $L(\Phi,\Theta)$:

$$R(\Phi,\Theta)=\sum_{i=1}^{k}\tau_i R_i(\Phi,\Theta), \qquad L(\Phi,\Theta)+R(\Phi,\Theta)\ \to\ \max_{\Phi,\Theta}, \tag{2.11}$$

$$\sum_{w\in W}\varphi_{wt}\in\{0,1\}, \qquad \varphi_{wt}\geqslant 0; \tag{2.12}$$

$$\sum_{t\in T}\theta_{td}\in\{0,1\}, \qquad \theta_{td}\geqslant 0. \tag{2.13}$$

где τ_i — неотрицательные *коэффициенты регуляризации*. Оптимизация взвешенной суммы критериев является широко распространённым приёмом в многокритериальной оптимизации. Преобразование вектора критериев в один скалярный критерий называется *скаляризацией*.

Модель вероятностного латентного семантического анализа PLSA соответствует частному случаю, когда регуляризатор отсутствует, $R(\Phi,\Theta)=0$.

В модели PLSA увеличение числа тем может приводить только к росту правдоподобия модели. Для регуляризованной модели это не обязательно так. Поэтому ограничения-равенства (2.12), (2.13) записаны с вариативной правой частью, предусматривающей возможность обнуления столбцов матриц Φ и Θ. Рассматриваются одновременно $2^{|T|}2^{|D|}$ задач, и из них выбирается та, для которой выполняются необходимые условия экстремума при заданных коэффициентах регуляризации.

Если $\varphi_t=0$, то тема t исключается из тематической модели. Таким образом, в постановку задачи закладывается возможность определять оптимальное число тем, при условии, что исходно было задано избыточное число тем. Заметим, что к удалению темы t может приводить не только обнуление t-го столбца матрицы Φ, но и обнуление t-й строки матрицы Θ.

Если $\theta_d=0$, то документ d фактически исключается из коллекции. Регуляризованная модель может отказываться определять тематику документа d, если он слишком короткий или если он не релевантен тематике коллекции.

В байесовских методах обучения тематических моделей [17, 14, 11] регуляризатор $R(\Phi,\Theta)$ интерпретируется как логарифм априорного распределения, а оптимизационная задача (2.11) соответствует принципу максимума апостериорной вероятности. В АРТМ регуляризатор не обязан иметь вероятностную интерпретацию.

Введём оператор неотрицательного нормирования, который преобразует произвольный вектор $(x_i)_{i\in I}$ в вектор вероятностей $(p_i)_{i\in I}$ дискретного распределения путём обнуления отрицательных элементов с последующей нормировкой:

$$p_i=\operatorname*{norm}_{i\in I} x_i=\frac{\max\{x_i,0\}}{\sum\limits_{j\in I}\max\{x_j,0\}}, \text{ для всех } i\in I.$$

Если $x_i\leqslant 0$ для всех $i\in I$, то результатом norm является нулевой вектор.

Теорема 2.1. *Пусть функция $R(\Phi,\Theta)$ непрерывно дифференцируема. Точка (Φ,Θ) локального экстремума задачи (2.11)–(2.13) удовлетворяет системе уравнений со*

вспомогательными переменными $p_{tdw} = p(t \mid d, w)$:

$$p_{tdw} = \underset{t \in T}{\operatorname{norm}}\big(\varphi_{wt}\theta_{td}\big); \tag{2.14}$$

$$\varphi_{wt} = \underset{w \in W}{\operatorname{norm}}\left(n_{wt} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}}\right); \qquad n_{wt} = \sum_{d \in D} n_{dw} p_{tdw}; \tag{2.15}$$

$$\theta_{td} = \underset{t \in T}{\operatorname{norm}}\left(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right); \qquad n_{td} = \sum_{w \in d} n_{dw} p_{tdw}. \tag{2.16}$$

Доказательство основано на применении условий Каруша–Куна–Таккера и приводится в [63, 62]. Данная теорема является следствием более общей теоремы 2.2, которая будет доказана ниже.

Решение системы уравнений (2.14)–(2.16) методом простых итераций приводит к регуляризованному ЕМ-алгоритму. Вычисление переменных p_{tdw} по формуле (2.14) называется Е-шагом, оценивание параметров $\varphi_{wt}, \theta_{td}$ по формулам (2.15) и (2.16) — М-шагом. Параметры $\varphi_{wt}, \theta_{td}$ можно инициализировать случайными значениями. Многочисленные разновидности ЕМ-алгоритма рассматриваются в [11, 3]. Ниже будет рассмотрен онлайновый ЕМ-алгоритм, который считается наиболее быстрым и хорошо распараллеливается. Именно он и реализован в библиотеке BigARTM.

§2.3 Латентное размещение Дирихле

Модель *латентного размещения Дирихле* (Latent Dirichlet Allocation, LDA) [17] основана на предположении, что столбцы θ_d и φ_t являются случайными векторами, порождаемыми распределениями Дирихле.

Априорное распределение Дирихле является сопряжённым к мультиномиальному распределению. Это упрощает применение методов байесовского обучения

В то же время, применение распределений Дирихле не имеет убедительных лингвистических обоснований. Его широкое распространение в тематическом моделировании объясняется скорее популярностью байесовского обучения, чем стремлением к адекватному моделированию значимых особенностей текстовых коллекций или явлений естественного языка.

В АРТМ модель LDA получает альтернативную не-вероятностную интерпретацию через сглаживающий регуляризатор [63]:

$$R(\Phi, \Theta) = \beta_0 \sum_{t,w} \beta_{wt} \ln \varphi_{wt} + \alpha_0 \sum_{d,t} \alpha_{td} \ln \theta_{td},$$

где β_0, α_0 — коэффициенты регуляризации. Максимизация $R(\Phi, \Theta)$ приводит к сближению вектор-столбцов φ_t с заданными векторами $\beta_t = (\beta_{wt}) \in \mathbb{R}^W$, а вектор-столбцов θ_d — с заданными векторами $\alpha_d = (\alpha_{td}) \in \mathbb{R}^T$. При этом векторы $\beta_0\beta_t$ и $\alpha_0\alpha_d$ соответствуют гиперпараметрам априорных распределений Дирихле.

Противоположная стратегия минимизации $R(\Phi, \Theta)$ приводит к эффекту разреживания, когда многие элементы матриц Φ и Θ обращаются в нуль. Она формализует *гипотезу разреженности* — естественное предположение, что каждый документ d и каждый термин w связан лишь с небольшим числом тем t.

В АРТМ сглаживающие и разреживающие регуляризаторы описываются одинаковыми формулами М-шага и отличаются только знаками коэффициентов β_{wt}, α_{dt}:

$$\varphi_{wt} = \underset{w\in W}{\operatorname{norm}}(n_{wt} + \beta_0\beta_{wt}), \qquad \theta_{td} = \underset{t\in T}{\operatorname{norm}}(n_{dt} + \alpha_0\alpha_{dt}),$$

при этом положительный коэффициент приводит к сглаживанию, отрицательный — к разреживанию соответствующего параметра φ_{wt} или θ_{td}.

В экспериментах разреженность матриц Φ и Θ может достигать более 90% практически без снижения правдоподобия модели [63], что подтверждает гипотезу разреженности. В [62] предлагается сглаживать небольшое число «фоновых» тем, чтобы собрать в них слова общей лексики, а основные «предметные» темы разреживать, чтобы в них сконцентрировались термины предметных областей.

В байесовском подходе нет единого описания сглаживания и разреживания, поскольку параметры распределений Дирихле β_{wt}, α_{td} могут быть только положительными. Известные попытки разреживания в байесовских моделях приводят к более громоздким конструкциям [52, 24, 69, 34, 20] из-за внутреннего противоречия между требованием разреженности и свойствами распределения Дирихле.

В АРТМ не используется байесовский вывод и сопряжённые распределения. Поэтому нет никаких оснований выделять сглаживающий регуляризатор Дирихле как основной. Далее мы будем рассматривать тематические модели без него, хотя многие из них исходно вводились на основе LDA.

§2.4 Мультимодальные тематические модели

Обобщим вероятностную тематическую модель (2.2) на случай конечного множества *модальностей*, каждая из которых имеет свой словарь — конечное множеством 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 $m(w)$. Первая модальность W^1 соответствует терминам (словам или словосочетаниям), остальные – различным типам метаданных.

Мультимодальное тематическое моделирование преследует одновременно две цели. С одной стороны, найти тематические профили документов $p(t\,|\,d)$, одинаково хорошо объясняющие появление токенов всех модальностей. С другой стороны, определить семантику и описать в терминах естественного языка элементы нетекстовых модальностей, таких, как элементы изображений или пользователи, рис. 2.

Коллекция документов рассматривается как случайная и независимая выборка троек (w_i, d_i, t_i), $i = 1, \ldots, n$ из дискретного распределения $p(w, d, t)$ на конечном вероятностном пространстве $W \times D \times T$. Обозначим через n_d число токенов всех модальностей в документе d, через n_{dw} — число вхождений токена w в документ d.

Примем гипотезу условной независимости $p(w\,|\,d, t) = p(w\,|\,t)$ и запишем тематическую модель модальности m, которая почти не отличается от модели (2.2):

$$p(w\,|\,d) = \sum_{t\in T} p(w\,|\,t)\,p(t\,|\,d) = \sum_{t\in T} \varphi_{wt}\theta_{td}, \qquad w \in W^m, \quad d \in D. \tag{2.17}$$

Параметры модели образуют матрицы $\Phi^m = \big(\varphi_{wt}\big)_{W^m\times T}$ и $\Theta = \big(\theta_{td}\big)_{T\times D}$. 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.

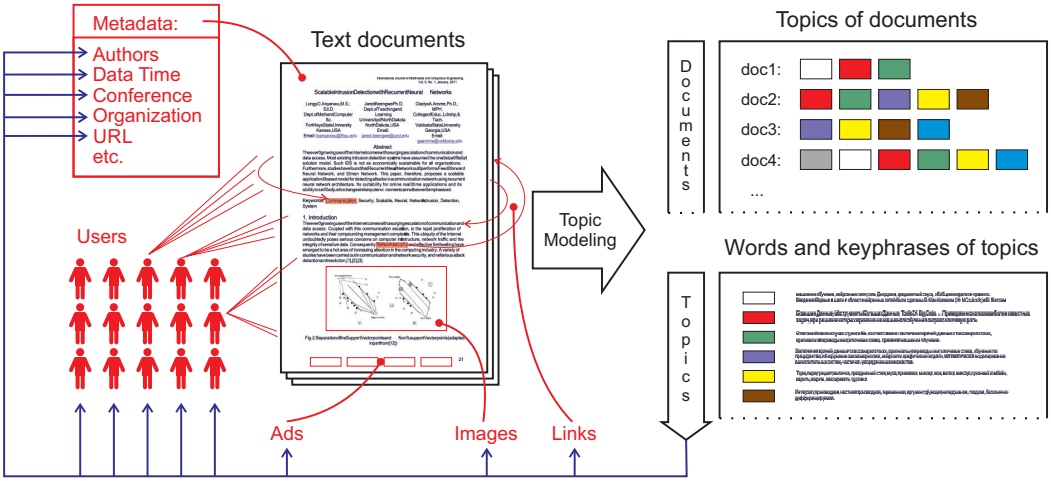


Рис. 2. Обычная тематическая модель определяет распределение тем в каждом документе $p(t\,|\,d)$ и распределение терминов в каждой теме $p(w\,|\,t)$. Мультимодальная модель определяет также распределения других модальностей в каждой теме. Примерами модальностей являются элементы метаописания документов (авторы, моменты времени, источники, рубрики и т. д.), цитаты и ссылки между документами, объекты на изображениях, рекламные баннеры, пользователи и т. д.

Запишем принцип максимума правдоподобия, представив логарифм правдоподобия в виде суммы по модальностям:

$$\ln\prod_{i=1}^{n} p(w_i\,|\,d_i) = \sum_{m\in M}\sum_{d\in D}\sum_{w\in W^m} n_{dw}\ln p(w\,|\,d) \to \max_{\Phi,\Theta}.$$

где n_{dw} — число вхождений токена w в документ d. Таким образом, логарифм правдоподобия мультимодальной тематической модели представляется суммой логарифмов правдоподобия тематических моделей отдельных модальностей:

$$L_m(\Phi^m,\Theta) = \sum_{d\in D}\sum_{w\in W^m} n_{dw}\ln p(w\,|\,d) \to \max_{\Phi^m,\Theta}. \tag{2.18}$$

Первое слагаемое в сумме $\sum_m L_m(\Phi^m,\Theta)$ относится к модальности терминов и совпадает с (2.5). Остальные слагаемые можно интерпретировать как регуляризаторы соответствующих модальностей. Введём в эту сумму коэффициенты регуляризации τ_m для каждой модальности m:

$$L(\Phi,\Theta) = \sum_{m\in M}\tau_m L_m(\Phi^m,\Theta) \to \max_{\Phi,\Theta},$$

что эквивалентно домножению чисел n_{dw} на эти коэффициенты: $\tilde{n}_{dw} = \tau_{m(w)} n_{dw}$. Коэффициенты τ_m позволяют сбалансировать модальности с учётом их важности и встречаемости в документах. Для модальности терминов будем полагать $\tau_1 = 1$.

С учётом остальных регуляризаторов и нормировки по каждой модальности задача построения тематической модели принимает следующий вид:

$$\sum_{d\in D}\sum_{w\in d}\tilde{n}_{dw}\ln\sum_{t\in T}\varphi_{wt}\theta_{td} + R(\Phi,\Theta) \to \max; \tag{2.19}$$

$$\sum_{w\in W^m}\varphi_{wt}\in\{0,1\}, \qquad \varphi_{wt}\geqslant 0; \tag{2.20}$$

$$\sum_{t\in T}\theta_{td}\in\{0,1\}, \qquad \theta_{td}\geqslant 0. \tag{2.21}$$

Теорема 2.2. *Пусть функция $R(\Phi,\Theta)$ непрерывно дифференцируема. Точка (Φ,Θ) локального экстремума задачи (2.19)–(2.21) удовлетворяет системе уравнений со вспомогательными переменными $p_{tdw}=p(t\,|\,d,w)$:*

$$p_{tdw} = \underset{t\in T}{\operatorname{norm}}\big(\varphi_{wt}\theta_{td}\big); \tag{2.22}$$

$$\varphi_{wt} = \underset{w\in W^m}{\operatorname{norm}}\left(n_{wt} + \varphi_{wt}\frac{\partial R}{\partial\varphi_{wt}}\right); \qquad n_{wt} = \sum_{d\in D}\tilde{n}_{dw}p_{tdw}; \tag{2.23}$$

$$\theta_{td} = \underset{t\in T}{\operatorname{norm}}\left(n_{td} + \theta_{td}\frac{\partial R}{\partial\theta_{td}}\right); \qquad n_{td} = \sum_{w\in W}\tilde{n}_{dw}p_{tdw}. \tag{2.24}$$

Теорема 2.1 является частным случаем теоремы 2.2 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 коэффициенты модальностей $\tau_{m(w)}$.

Тематическая модель классификации Dependency LDA [51] аналогична модели (2.17) и имеет две модальности — термины и классы. Каждому документу d соответствует подмножество меток классов $C_d\subset C$. Тематические модели превосходят обычные методы классификации на больших текстовых коллекциях с большим числом несбалансированных, пересекающихся, взаимозависимых классов [51]. *Несбалансированность* означает, что классы могут содержать как очень малое, так и очень большое число документов. В случае *пересекающихся* классов документ может относиться как к одному классу, так и к очень большому числу классов. *Взаимозависимые* классы имеют схожие множества характерных терминов, и при классификации документа могут вступать в конкуренцию.

Тематическая модель CI-LDA (Conditionally Independent LDA) [41] аналогична модели (2.17) и имеет две модальности — термины и именованные сущности. В названии подчёркивается, что модель опирается на гипотезу условной независимости двух модальностей.

Тематическая модель цитирования документов LDA-post [23] аналогична модели (2.17) с двумя модальностями. Первой модальностью являются термины. Второй модальностью являются документы, C_d — это множество документов, процитированных в документе d. При этом число ненулевых элементов в строке матрицы Φ^2, соответствующей документу $c\in D$, интерпретируется как число тем, на которые документ c оказывает существенное влияние.

§2.5 Онлайновый EM-алгоритм

Вычисление параметров модели φ_{wt}, θ_{td} на M-шаге требует однократного прохода всей коллекции в цикле по всем документам $d \in D$ и всем токенам каждого документа $w \in d$. Переменные p_{tdw} можно вычислять внутри этого цикла, непосредственно в тот момент, когда они понадобятся. Таким образом, E-шаг встраивается внутрь M-шага, что позволяет избежать хранения трёхмерной матрицы p_{tdw} без дополнительных вычислительных затрат. Существует множество версий EM-алгоритма, различающихся частотой обновления параметров модели φ_{wt} и θ_{td} по переменным n_{wt} и n_{td}. Частые обновления повышают скорость сходимости и почти не влияют на значение правдоподобия в конце итераций [3].

Для обработки больших коллекций лучше всего подходят *онлайновые алгоритмы* Online PLSA [12] и Online LDA [26]. Последний реализован в библиотеке онлайновых методов машинного обучения Vowpal Wabbit и считается одной из самых эффективных реализаций вероятностного тематического моделирования. Онлайновые алгоритмы основаны на следующей стратегии обновлений. Итерации θ_{td} со встроенным E-шагом производятся при фиксированной матрице Φ для каждого документа d до сходимости. На последней итерации документа производится накопительное обновление переменных n_{wt}. Обновления матрицы Φ по переменным n_{wt} происходят по окончании обработки документа, либо ещё реже — по окончании обработки пакета документов. На больших коллекциях матрица Φ обычно сходится после обработки относительно небольшой части документов. В результате даже одного прохода по коллекции бывает достаточно для построения модели. Это позволяет применять онлайновые алгоритмы для анализа новостных потоков.

Алгоритм 2.3 показывает организацию вычислительного процесса для коллекции D, разбитой на пакеты документов D_b, $b = 1, \ldots, B$. Обработка каждого пакета выполняется Алгоритмом 2.4. В моменты синхронизации происходит объединение накопленных обновлений $\tilde{n}_{wt}$ в матрице Φ, см. шаги 5–8. Понижая частоту синхронизации, можно добиться наивысшей скорости обработки практически без ущерба для качества модели.

Коэффициент дисконтирования ρ обычно полагается равным единице. Его можно понижать в тех случаях, когда коллекция обрабатывается за один проход, и темы должны определяться по наиболее свежим документам. Коэффициент ρ задаёт экспоненциальное уменьшение весов ранее обработанных пакетов.

§2.6 Параллельный онлайновый EM-алгоритм

При разработке параллельной архитектуры BigARTM учитывался опыт известных параллельных реализаций алгоритмов оптимизации тематических моделей.

Алгоритм Approximate Distributed LDA (AD-LDA) [39]. Коллекция распределяется по процессорам примерно в равных пропорциях, по окончании обработки данных всеми процессорами производится синхронизация, в результате которой обновляется глобальная матрица Φ. Недостатком этой реализации является то, что время, затрачиваемое на одну итерацию работы алгоритма, определяется самым медленным из процессоров. Сеть во время итераций простаивает, а во время синхронизаций пере-

Алгоритм 2.3. Онлайновый EM-алгоритм для мультимодальной APTM

Вход: коллекция D_b, коэффициент дисконтирования $\rho \in (0, 1]$;
Выход: матрица Φ;

инициализировать φ_{wt} для всех $w \in W,\ t \in T$;
$n_{wt} := 0,\ \tilde{n}_{wt} := 0$ для всех $w \in W,\ t \in T$;
для всех пакетов D_b, $b = 1, \ldots, B$
    $(\tilde{n}_{wt}) := (\tilde{n}_{wt}) + \mathsf{ProcessBatch}(D_b, \Phi)$;
    если пора выполнить синхронизацию, **то**
        $n_{wt} := \rho n_{wt} + \tilde{n}_{dw}$ для всех $w \in W,\ t \in T$;
        $\varphi_{wt} := \underset{w \in W^m}{\operatorname{norm}}\left(n_{wt} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}}\right)$ для всех $w \in W^m,\ m = 1, \ldots, M,\ t \in T$;
        $\tilde{n}_{wt} := 0$ для всех $w \in W,\ t \in T$;

Алгоритм 2.4. $\mathsf{ProcessBatch}(D_b, \Phi)$

Вход: пакет D_b, матрица $\Phi = (\varphi_{wt})$;
Выход: матрица $(\tilde{n}_{wt})$;

$\tilde{n}_{wt} := 0$ для всех $w \in W,\ t \in T$;
для всех $d \in D_b$
    инициализировать $\theta_{td} := \frac{1}{|T|}$ для всех $t \in T$;
    повторять
        $p_{tdw} := \underset{t \in T}{\operatorname{norm}}\left(\varphi_{wt}\theta_{td}\right)$ для всех $w \in d,\ t \in T$;
        $n_{td} := \sum_{w \in d} \tau_{m(w)} n_{dw} p_{tdw}$ для всех $t \in T$;
        $\theta_{td} := \underset{t \in T}{\operatorname{norm}}\left(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right)$ для всех $t \in T$;
    пока θ_d не сойдётся;
    $\tilde{n}_{wt} := \tilde{n}_{wt} + \tau_{m(w)} n_{dw} p_{tdw}$ для всех $w \in d,\ t \in T$;

гружена. Кроме того, для каждого ядра хранится копия матрицы Φ, что приводит к необоснованному расходу памяти.

Алгоритм Y!LDA [55]. Коллекция распределяется по узлам, на каждом узле создаётся несколько потоков, занимающихся обработкой документов, и один поток, производящий обновление глобальных счётчиков n_{wt}. К этому потоку обработчики обращаются асинхронно по мере готовности новых обновлений. Параллелизм, основанный на многопоточности, позволяет хранить одну копию глобальных счётчиков для всех ядер на каждом узле. Синхронизация состояний всех узлов организована на основе архитектуры классной доски. Суть её в том, что глобальное состояние хранится в некоторой общей для всех узлов памяти, и поток синхронизации каждого узла асинхронно обращается к ней и производит обновление. Система производит параллельную обработку документов, эффективно используя имеющиеся вычислительные и сетевые ресурсы.

Алгоритм Mr.LDA [77] основан на технологиях MapReduce и Hadoop. На шаге Map производится оптимизация параметров, связанных с документами, на шаге Reduce — с темами. Технология MapReduce обеспечивает хорошую отказоустойчивость.

Библиотека Gensim предоставляет две реализации LDA на языке Python: LdaModel и LdaMulticore [65]. Первая из них почти в точности повторяет Online LDA из [26] и не является ни параллельной, ни распределённой. Тем не менее, LdaModel работает достаточно эффективно за счёт использования матричных вычислений в библиотеке NumPy. LdaMulticore является параллельным, архитектурно он имеет много общего с Y!LDA.

Библиотека BigARTM разрабатывается исходя из требований асинхронной обработки данных, минимизации используемого объёма оперативной памяти, масштабируемости при увеличении количества ядер на узле, кроссплатформенности, возможности быстрой установки и использования на одной машине.

В результате анализа имеющихся параллельных реализаций было принято решение не использовать MapReduce. Во-первых, обработка данных в рамках этого подхода является синхронной, во-вторых, такое решение плохо подходит для использования на одном компьютере. MapReduce прежде всего ориентирован на применение простых операций к большим объёмам данных. Для реализации итеративных алгоритмов оптимизации с частичным обновлением большой глобальной матрицы Φ ему не хватает гибкости.

Для организации параллельной обработки данных на одном узле в BigARTM используется многопоточный параллелизм в пределах одного процесса. Это позволяет получить хорошую скорость обработки и хранить общую матрицу Φ для узла, а не для каждого ядра. Кроме того, такое решение обеспечивает возможность асинхронной работы с данными.

Библиотека BigARTM предназначена прежде всего для эффективной онлайновой параллельной обработки в пределах одной машины. Для достижения независимости объёма используемой оперативной памяти от размера обрабатываемой коллекции вся коллекция D делится на пакеты D_b, $b = 1, \ldots, B$, которые сохраняются на диск в отдельных файлах. В каждый момент времени в памяти находится только часть из них. Вся матрица Φ никогда не хранится. В каждый момент времени в памяти содержится только её фрагмент, соответствующий обрабатываемым документам, и после окончания обработки они удаляются.

Организация параллельной обработки в BigARTM, как в Y!LDA и Gensim, основана на многопоточности. Существует множество обработчиков (Processor) и один поток слияния (Merger), рис. 3.

Обработчики производят параллельную обработку пакетов документов, вычисляют θ_{td} и накапливают счётчики n_{wt} для обновления матрицы Φ. Поток слияния асинхронно получает значения счётчиков от обработчиков n_{wt} и обновляет матрицу Φ. Для обработчиков существует очередь заданий, в которых содержатся подгруженные библиотекой в память пакеты. Также существует очередь слияния, в которую обработчики отправляют по мере готовности обновления матрицы Φ для потока слияния. Все данные хранятся в памяти статично и не копируются, перемещения производятся с помощью указателей. В каждый момент времени Merger хранит две

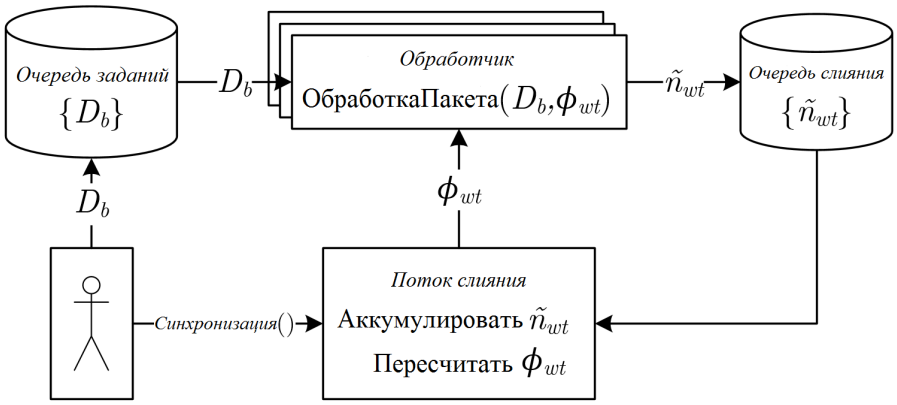


Рис. 3. Организация параллельной обработки в BigARTM.

копии матрицы Φ, базовую и активную. Первая из них доступна для чтения обработчикам, вторая доступна для записи потоку слияния. Каждый обработчик перед началом работы захватывает указатель на текущую активную матрицу Φ и использует её для вычисления параметров θ_{td} своего пакета документов. Поток слияния, получая обновления из очереди слияния, обновляет базовую матрицу Φ. Как только все текущие обновления завершаются и обработчики отпускают указатели активной матрицы, поток слияния делает базовую матрицу активной и создаёт новую базовую матрицу. Обновление матрицы Φ (синхронизацию) можно инициировать в любой момент из пользовательского кода.

3 Библиотека регуляризаторов

В данном разделе рассматриваются регуляризаторы, которые уже реализованы или планируются для реализации в библиотеке BigARTM. Другие регуляризаторы для APTM можно найти в [63, 62].

Дивергенция Кульбака–Лейблера или *KL-дивергенция* будет активно использоваться при построении регуляризаторов. KL-дивергенция — это несимметричная функция расстояния между дискретными распределениями $P = (p_i)_{i=1}^n$ и $Q = (q_i)_{i=1}^n$:

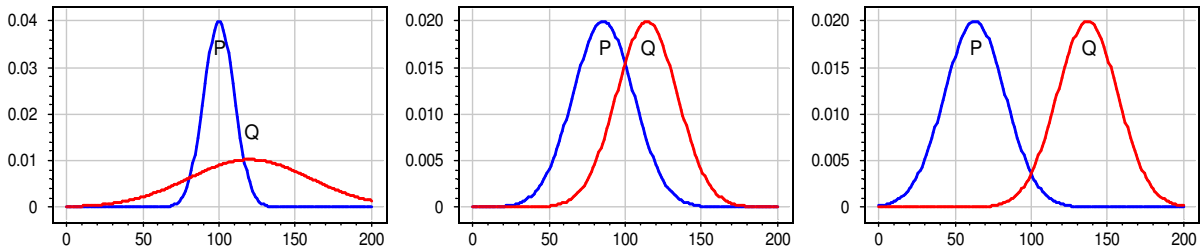
$$\mathrm{KL}(P\|Q) \equiv \mathrm{KL}_i(p_i\|q_i) = \sum_{i=1}^{n} p_i \ln \frac{p_i}{q_i}.$$

Предполагается, что $p_i > 0$ и $q_i > 0$. KL-дивергенция является не вполне адекватной функцией расстояния в случае, когда у распределений P и Q не совпадают носители $\Omega_P = \{i\colon p_i > 0\}$ и $\Omega_Q = \{i\colon q_i > 0\}$.

Перечислим наиболее важные свойства KL-дивергенции.

1. KL-дивергенция неотрицательна. Если $\Omega_P = \Omega_Q$, то KL-дивергенция равна нулю тогда и только тогда, когда распределения совпадают, $p_i \equiv q_i$.

2. KL-дивергенция является мерой вложенности двух распределений. Если $\mathrm{KL}(P\|Q) < \mathrm{KL}(Q\|P)$, то распределение P сильнее вложено в Q, чем Q в P, см. рис. 4.



$\mathrm{KL}(P\|Q) = 0.442$ $\mathrm{KL}(Q\|P) = 2.966$ | $\mathrm{KL}(P\|Q) = 0.444$ $\mathrm{KL}(Q\|P) = 0.444$ | $\mathrm{KL}(P\|Q) = 2.969$ $\mathrm{KL}(Q\|P) = 2.969$

Рис. 4. Дивергенция Кульбака–Лейблера $\mathrm{KL}(P\|Q)$ является несимметричной мерой вложенности распределения $P = (p_i)_{i=1}^n$ в распределение $Q = (q_i)_{i=1}^n$. Вложенность P в Q приблизительно одинакова на левом и среднем графиках, вложенность Q в P — на левом и правом графиках.

3. Если P — эмпирическое распределение, а $Q(\alpha)$ — параметрическое семейство (модель) распределений, то минимизация KL-дивергенции эквивалентна максимизации правдоподобия:

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

Максимизация правдоподобия (2.5) эквивалентна минимизации взвешенной суммы дивергенций Кульбака–Лейблера между эмпирическими распределениями $\hat{p}(w \,|\, d) = \frac{n_{dw}}{n_d}$ и модельными $p(w \,|\, d)$, по всем документам d из D:

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

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

§3.1 Сглаживание, разреживание и частичное обучение

Сглаживание. Потребуем, чтобы столбцы φ_t и θ_d были близки к заданным распределениям $\beta_t = (\beta_{wt})_{w\in W^m}$ и $\alpha_d = (\alpha_{td})_{t\in T}$ в смысле дивергенции Кульбака–Лейблера:

$$\sum_{t\in T} \mathrm{KL}_w(\beta_{wt}\|\varphi_{wt}) \to \min_{\Phi^m}, \qquad \sum_{d\in D} \mathrm{KL}_t(\alpha_{td}\|\theta_{td}) \to \min_{\Theta}.$$

Складывая два критерия с коэффициентами β_0, α_0 и удаляя из суммы константы, получим регуляризатор

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

Применение общих формул (2.15) и (2.16) даёт выражение для M-шага:

$$\varphi_{wt} = \underset{w\in W^m}{\mathrm{norm}}\big(n_{wt} + \beta_0\beta_{wt}\big); \qquad \theta_{td} = \underset{t\in T}{\mathrm{norm}}\big(n_{td} + \alpha_0\alpha_{td}\big).$$

Сглаживающий регуляризатор эквивалентен предположению, что столбцы матриц Φ^m и Θ порождаются априорными распределениями Дирихле с гиперпараметрами $\beta_0\beta_t$ и $\alpha_0\alpha_d$. В модели *латентного размещения Дирихле* LDA [17] гиперпараметры могут быть только положительными.

Векторы β_t и α_d обычно берут одинаковыми для всех столбцов матриц Φ^m и Θ. В большинстве исследований используются симметричные распределения Дирихле (с равными значениями всех координат в векторах β_t и α_d), хотя известно, что оптимизация гиперпараметров улучшает качество модели [68].

Разреживание. Недостатком сглаживающего регуляризатора является его противоречие с *гипотезой разреженности*. Естественно предполагать, что каждый документ d и каждый токен w связан лишь с небольшим числом тем t. В таком случае значительная часть вероятностей φ_{wt} и θ_{td} должны быть равны нулю.

Чем сильнее разрежено распределение, тем ниже его энтропия. Максимальной энтропией обладает равномерное распределение. Идея разреживания состоит в том, чтобы максимизировать дивергенции $\mathrm{KL}_w\big(\frac{1}{|W|}\|\varphi_{wt}\big)$ и $\mathrm{KL}_t\big(\frac{1}{|T|}\|\theta_{td}\big)$ между искомыми распределениями и равномерными. Обобщая эту идею, зададим вместо равномерных распределений произвольные распределения $\beta_t = (\beta_{wt})_{w\in W^m}$ и $\alpha_d = (\alpha_{td})_{t\in T}$. В таком случае разреживание оказывается прямой противоположностью сглаживанию:

$$R(\Phi,\Theta) = -\beta_0 \sum_{t\in T}\sum_{w\in W^m} \beta_{wt}\ln\varphi_{wt} - \alpha_0 \sum_{d\in D}\sum_{t\in T} \alpha_{td}\ln\theta_{td} \to \max;$$

$$\varphi_{wt} = \underset{w\in W^m}{\mathrm{norm}}\big(n_{wt} - \beta_0\beta_{wt}\big); \qquad \theta_{td} = \underset{t\in T}{\mathrm{norm}}\big(n_{td} - \alpha_0\alpha_{td}\big).$$

В BigARTM регуляризаторы разреживания и сглаживания объединены в один регуляризатор и отличаются только знаками элементов матриц (β_{wt}) и (α_{td}). Это позволяет комбинировать сглаживающие и разреживающие воздействия, а также задавать область их действия подмножествами строк и столбцов.

Заметим, что априорные распределения, соответствующие такому разреживающему регуляризатору, в байесовском подходе никогда не рассматривались.

Идея энтропийного разреживания была предложена в динамической тематической модели PLSA для обработки видеопотоков [60]. В данной задаче документами являются короткие видеофрагменты, терминами — признаки на изображениях, темами — появление определённого объекта в течение определённого времени, например, проезд автомобиля. Сильно разреженные распределения требовались для описания тем с кратким «временем жизни».

Частичное обучение. В ходе анализа построенной тематической модели эксперты (асессоры) могут отмечать релевантные и нерелевантные токены в темах и темы в документах. Эти данные могут затем использоваться как для количественного оценивания и сравнения моделей, так и для дальнейшего улучшения её качества.

Регуляризация по размеченным данным позволяет зафиксировать интерпретации тем и повышает устойчивость модели. Обычно разметка делается для небольшой доли документов и тем, поэтому задача использования разметки относится к области *частичного обучения* (semi-supervised learning).

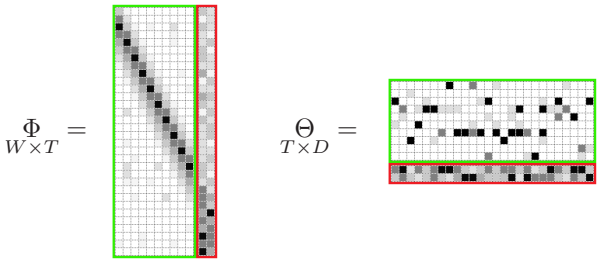


Рис. 5. Структура разреженности матриц Φ и Θ с предметными и фоновыми темами.

Пусть $D_0 \subset D$ — подмножество размеченных документов, и для каждого документа $d \in D_0$ задано подмножество релевантных тем $T_d \subset T$, к которым он относится, и подмножество нерелевантных тем $\bar{T}_d \subset T$, к которым он не относится.

Пусть $T_0 \subset T$ — подмножество размеченных тем, и для каждой темы $t \in T_0$ задано подмножество релевантных токенов $W_t \subset W$, которые к ней относятся, и подмножество нерелевантных токенов $\bar{W}_t \subset W$, которые к ней не относятся.

Регуляризатор *частичного обучения по релевантности* является частным случаем сглаживающего регуляризатора при

$$\beta_{wt} = \tfrac{1}{|W_t|}[w \in W_t][t \in T_0],$$
$$\alpha_{td} = \tfrac{1}{|T_d|}[t \in T_d][d \in D_0].$$

Регуляризатор *частичного обучения по нерелевантности* является частным случаем разреживающего регуляризатора при

$$\beta_{wt} = -\tfrac{1}{|W_t|}[w \in \bar{W}_t][t \in T_0],$$
$$\alpha_{td} = -\tfrac{1}{|T_d|}[t \in \bar{T}_d][d \in D_0].$$

При достаточно больших значениях коэффициентов регуляризации β_0 и α_0 он обнуляет вероятности нерелевантных тем в документах и токенов в темах.

Выделение предметных и фоновых тем. Чтобы тема была интерпретируемой, она должна содержать лексическое ядро — множество слов, характерных для определённой предметной области, которые часто употребляются рядом в документах, с большой вероятностью употребляются в данной теме и практически не употребляются в других темах. Отсюда следует, что из бесконечного множества стохастических матричных разложений $F \approx \Phi\Theta$ нас больше всего интересуют те, в которых матрицы Φ и Θ обладают структурой разреженности, примерно показанной на рис. 5.

Множество тем разбивается на два подмножества, $T = S \sqcup B$: предметные темы S и фоновые темы B.

Предметные темы $t \in S$ содержат термины предметных областей. Их распределения $p(w \mid t)$ разрежены и существенно различны (декоррелированы). Распределения $p(d \mid t)$ 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 $p(w \mid t)$ и $p(d \mid t)$ сглажены, так как эти слова присутствуют в большинстве документов. Тематическую модель с фоновыми темами можно рассматривать как обобщение робастных моделей [19, 48], в которых использовалось только одно фоновое распределение.

§3.2 Декоррелирование

Тематическая модель тем полезнее, чем более различные темы она находит. Это предположение приводит к дополнительному требованию увеличивать различность тем. Можно по-разному формализовать понятие различности тем как дискретных распределений $\varphi_{wt} = p(w \mid t)$ или нормированных векторов $\varphi_w = (\varphi_{wt})_{t \in T}$. Остановимся на естественной мере различности — ковариации:

$$R(\Phi, \Theta) = -\frac{\tau}{2} \sum_{t \in T} \sum_{s \in T \setminus t} \operatorname{cov}(\varphi_t, \varphi_s) \to \max, \qquad \operatorname{cov}(\varphi_t, \varphi_s) = \sum_{w \in W} \varphi_{wt} \varphi_{ws}.$$

Этот критерий не зависит от Θ, поэтому для θ_{td} формулы М-шага не меняются. Формула для φ_{wt}, согласно (2.15), принимает вид

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

Смысл этой формулы в том, что условные вероятности $\varphi_{wt} = p(w \mid t)$ постепенно уменьшаются для тех терминов w, которые имеют бóльшие значения вероятности φ_{ws} в других темах. Вероятности φ_{wt} наиболее значимых тем токена w в ходе итераций становятся ещё больше. Вероятности менее значимых тем постепенно уменьшаются и могут обращаться в нуль. Таким образом, данный регуляризатор также является разреживающим. Однако минимизация ковариаций не так агрессивно обнуляет строки матрицы Φ, соответствующие редким терминам, как разреживающие регуляризаторы. Кроме того, регуляризатор декоррелирования тем обладает дополнительным полезным свойством группировать слова общей лексики в отдельные темы [57]. Эксперименты с комбинированием регуляризаторов сглаживания, разреживания и декоррелирования в АРТМ полностью подтверждают это наблюдение [4, 62, 63].

§3.3 Отбор тем

Для удаления незначимых тем из тематической модели в [63] был предложен *регуляризатор отбора тем*, основанный на идее энтропийного разреживания распределения $p(t)$, которое легко выражается через параметры тематической модели:

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

Подставим этот регуляризатор в формулу М-шага (2.16):

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

Немного модифицируем эту формулу, заменив θ_{td} в правой части равенства несмещённой оценкой $\frac{n_{td}}{n_d}$:

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

В этом случае 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 $|T|$. Затем в ходе итераций число нулевых строк матрицы Θ постепенное увеличивается.

Множитель $\frac{n}{|T|}$ является естественной нормировкой, благодаря которой коэффициент регуляризации τ может выбираться из отрезка $[0,1]$. При $\tau=1$ на каждой итерации обнуляются все темы, в которых число токенов меньше среднего, $n_t < \frac{n}{|T|}$. Такое воздействие регуляризатора на модель представляется чрезмерно сильным, поэтому значение τ рекомендуется выбирать меньше единицы.

Заметим, что подход АРТМ к отбору тем намного проще непараметрических байесовских моделей, основанных на иерархических процессах Дирихле (Hierarchical Dirichlet Process, HDP) [58] или процессах китайского ресторана (Chinese Restaurant Process, CRP) [15]. Кроме того, он точнее определяет истинное число тем в экспериментах на полусинтетических данных и гораздо устойчивее определяет число тем на реальных данных [64]. Также замечено, что регуляризатор отбора тем постепенно выводит из модели вторичные темы, которые являются либо линейными комбинациями других тем, либо результатом расщепления других тем.

§3.4 Когерентность

Тема называется *когерентной* (согласованной), если термины, наиболее частые в данной теме, неслучайно часто совместно встречаются рядом в документах коллекции [42, 43]. Когерентность может оцениваться как по самой коллекции D [38], так и по сторонней коллекции, например, по Википедии [40]. Средняя когерентность тем считается хорошей мерой интерпретируемости тематической модели [43].

Пусть заданы оценки совместной встречаемости $C_{wv} = \hat{p}(w \,|\, v)$ для пар терминов $(w,v) \in W^1 \times W^1$. Обычно C_{wv} оценивают как долю документов, содержащих термин v, в которых термин w встречается не далее чем через 10 слов от v. Сам термин не учитывается, то есть полагают $C_{vv} = 0$. Эти оценки вычисляются на этапе предварительной обработки коллекции D.

Запишем оценку условной вероятности $\hat{p}(w \,|\, t)$ через условные вероятности $\varphi_{vt} = p(v \,|\, t)$ всех терминов v, когерентных с w:

$$\hat{p}(w \,|\, t) = \sum_{v \in W^1} \hat{p}(w \,|\, v) p(v \,|\, t) = \sum_{v \in W^1} C_{wv} \varphi_{vt} = \sum_{v \in W^1} \frac{C_{wv} n_{vt}}{n_t}.$$

Введём регуляризатор, требующий, чтобы оценка $\hat{p}(w \,|\, t)$ была согласована с тематической моделью, то есть близка к φ_{wt} по KL-дивергенции:

$$R(\Phi) = \tau \sum_{t \in T} n_t \sum_{w \in W^1} \hat{p}(w \,|\, t) \ln \varphi_{wt} \to \max.$$

Формула М-шага, согласно (2.15), принимает вид

$$\varphi_{wt} = \underset{w \in W^1}{\operatorname{norm}} \left(n_{wt} + \tau \sum_{v \in W^1} C_{wv} n_{vt} \right).$$

Эта же формула предлагалась в [38] для модели LDA и алгоритма сэмплирования Гиббса, с более сложным обоснованием через обобщённую урновую схему Пойя, и более сложной эвристической оценкой C_{wv}.

В работе [40] предлагалось использовать другой регуляризатор:

$$R(\Phi) = \tau \sum_{t \in T} \ln \sum_{u,v \in W^1} C_{uv} \varphi_{ut} \varphi_{vt} \to \max,$$

и другую оценку совместной встречаемости $C_{uv} = N_{uv}\big[\mathrm{PMI}(u,v) > 0\big]$, где N_{uv} — число документов, в которых термины u, v хотя бы один раз встречаются рядом (не далее, чем через 10 слов), $\mathrm{PMI}(u,v) = \ln \frac{|D| N_{uv}}{N_u N_v}$ — поточечная взаимная информация (pointwise mutual information), N_u — число документов, в которых термин u встречается хотя бы один раз.

Таким образом, в литературе пока отсутствует единый подход к оптимизации когерентности. Известные подходы легко формализуются в рамках ARTM и не требуют введения априорных распределений Дирихле.

§3.5 Сглаживание и разреживание во времени

TODO: Никита Дойков

§3.6 Классификация

TODO: AUC и ковариационный регуляризатор.

§3.7 Регрессия

TODO: Евгений Соколов

4 Библиотека метрик качества

Количественное оценивание тематических моделей является нетривиальной проблемой. В отличие от задач классификации или регрессии здесь нет чёткого понятия «ошибки» или «потери». Критерии качества кластеризации, типа средних внутрикластерных или межкластерных расстояний, плохо подходят для оценивания «мягкой» совместной кластеризации документов и терминов.

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 конечных пользователей. Для этого приходится собирать дополнительные данные, например, оценки асессоров.

§4.1 Перплексия

Наиболее распространённым внутренним критерием является *перплексия* (perplexity), используемая для оценивания моделей языка в вычислительной лингвистике. Это мера несоответствия или «удивлённости» модели $p(w \,|\, d)$ токенам w,

наблюдаемым в документах d коллекции D. Она определяется через логарифм правдоподобия (2.18), отдельно для каждой модальности:

$$\mathscr{P}_m(D;p) = \exp\left(-\frac{1}{n_m} L_m(\Phi,\Theta)\right) = \exp\left(-\frac{1}{n_m}\sum_{d\in D}\sum_{w\in W^m} n_{dw}\ln p(w\,|\,d)\right), \qquad (4.1)$$

где $n_m = \sum\limits_{d\in D}\sum\limits_{w\in W^m} n_{dw}$ — длина коллекции по m-й модальности.

Чем меньше величина перплексии, тем лучше модель p предсказывает появление токенов w в документах d коллекции D.

Перплексия имеет следующую интерпретацию. Если термины 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. Чем сильнее модель $p(w)$ отличается от генерирующего распределения, тем больше перплексия. В случае условных вероятностей $p(w\,|\,d)$ интерпретация немного другая: если каждый документ генерируется из V равновероятных терминов (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 $\mathscr{P}_m(D;p_D)$ является оптимистично смещённой (заниженной) характеристикой качества модели из-за эффекта переобучения. Обобщающую способность тематических моделей принято оценивать *перплексией контрольной выборки* (hold-out perplexity) $\mathscr{P}_m(D';p_D)$. Обычно коллекцию разделяют на обучающую и контрольную случайным образом в пропорции $9:1$ [17].

Недостатком контрольный 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 [17]. В [2, 48, 3] были предложены *робастные тематические модели*, описывающие редкие слова специальным «фоновым» распределением. Перплексия робастных вариантов PLSA и LDA оказалась существенно меньшей и практически одинаковой.

§4.2 Когерентность

Тема называется *когерентной* (согласованной), если термины, наиболее частые в данной теме, неслучайно часто совместно встречаются рядом в документах коллекции [42, 43]. Когерентность может оцениваться по сторонней коллекции (например, по Википедии) [40], либо по той же коллекции, по которой строится модель [38].

Для оценивания когерентности в [42, 43] использовалась *поточечная взаимная информация* (pointwise mutual information, PMI):

$$\mathrm{PMI}(t) = \sum_{i=1}^{k-1}\sum_{j=i}^{k} \log\frac{N(w_i,w_j)}{N(w_i)N(w_j)},$$

где w_i — i-й термин в порядке убывания φ_{wt}, $N(w)$ — число документов, в которых термин w встречается хотя бы один раз, $N(w, w')$ — число документов, в которых термины w, w' хотя бы один раз встречаются рядом (в окне заданной ширины $h = 10$). Число k обычно полагается равным 10.

Средняя когерентность тем считается хорошей мерой интерпретируемости тематической модели [43]. Среди внутренних критериев она имеет самую высокую корреляцию с экспертными оценками интерпретируемости.

§4.3 Тест условной независимости

TODO: Юлия Молчанова

§4.4 Разреженность

Разреженность модели измеряется долей $\mathscr{S}_\Phi$ и $\mathscr{S}_\Theta$ 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.

§4.5 Характеристики ядер тем

Предполагается, что интерпретируемая тема должна содержать лексическое ядро — множество слов, существенно отличающих её от остальных тем.

Формально определим ядро W_t темы t как множество терминов, которые имеют высокую условную вероятность $p(t \mid w) = \varphi_{wt}\frac{n_t}{n_w}$ для данной темы:

$$W_t = \big\{w \in W \;\big|\; p(t \mid w) > 0.25\big\}.$$

По ядру определим три показателя интерпретируемости темы t:
$\mathrm{pur}_t = \sum\limits_{w \in W_t} p(w \mid t)$ — *чистота темы* (чем выше, тем лучше);
$\mathrm{con}_t = \frac{1}{|W_t|}\sum\limits_{w \in W_t} p(t \mid w)$ — *контрастность темы* (чем выше, тем лучше);
$\mathrm{ker}_t = |W_t|$ — *размер ядра* (ориентировочный оптимум $\frac{|W|}{|T|}$).

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

§4.6 Доля фоновых слов

Доля фоновых слов во всей коллекции

$$\mathscr{B} = \frac{1}{n}\sum_{d \in D}\sum_{w \in d}\sum_{t \in B} n_{dw} p(t \mid d, w)$$

принимает значения от 0 до 1. Значения, близкие к 0, говорят о том, что модель не способна отделять слова общей лексики от специальной терминологии. Значения, близкие к 1, свидетельствуют о вырождении тематической компоненты модели, например, в результате чрезмерного разреживания.

§4.7 Качество тематического поиска

TODO: Марина Дударенко

5 Методы инициализации

§5.1 Контекстная кластеризация документов

TODO: Алексей Гринчук

§5.2 Кластеризация якорных слов

TODO: Андрей Шадриков

6 Стратегии регуляризации

§6.1 Относительные коэффициенты регуляризации

В библиотеке BigARTM используются *относительные коэффициенты регуляризации*, показывающие степень воздействия регуляризатора на параметры модели. Это позволяет локализовать область значений коэффициентов регуляризации в отрезке $[0, 1]$ и использовать стратегии регуляризации, слабо зависящие от размеров коллекции, числа тем, разреженности данных и других характеристик задачи.

Регуляризаторы, зависящие от матрицы Φ. Запишем формулу M-шага (2.15) с суммой регуляризаторов $\tau_i R_i$:

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

Определим *воздействие* r_{it} регуляризатора R_i на тему t и его *воздействие* r_i на всю матрицу Φ при $\tau_i = 1$:

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

Если $r_i = n$, то воздействие регуляризатора сопоставимо с объёмом коллекции $n = \sum_t \sum_w n_{wt}$. Введение отношения $\frac{n}{r_i}$ в формулу M-шага позволяет интерпретировать коэффициент регуляризации τ_i как *относительное воздействие* регуляризатора R_i на матрицу Φ:

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

Относительное воздействие τ_i показывает, во сколько раз действие регуляризатора на матрицу Φ сильнее самих данных. Слишком малые значения говорят о том,

что регуляризатор практические не используется. Слишком большие значения могут сигнализировать об избыточной регуляризации и возможном вырождении модели.

Если $r_{it} = n_t$, то воздействие регуляризатора на тему t сопоставимо с n_t — объёмом темы t в коллекции. Введение отношения $\frac{n_t}{r_{it}}$ в формулу М-шага позволяет интерпретировать коэффициент τ_i как *относительное воздействие* регуляризатора R_i на каждую тему:

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

Теперь абсолютное воздействие регуляризатора $\tau_i \frac{n_t}{r_{it}}$ на «малые темы» с низкими n_t становится меньше, на «большие темы» с высокими n_t, наоборот, увеличивается.

Возникает вопрос, в какой степени адаптировать регуляризаторы под каждую тему, и что лучше — чтобы абсолютные или относительные воздействия на темы были одинаковы? Для ответа на этот вопрос введём параметризацию:

$$\varphi_{wt} = \underset{w\in W}{\text{norm}}\left(n_{wt} + \sum_{i=1}^{k} \tau_i \left(\gamma_i \frac{n_t}{r_{it}} + (1-\gamma_i)\frac{n}{r_i}\right) \varphi_{wt} \frac{\partial R_i}{\partial \varphi_{wt}}\right).$$

Параметр γ_i назовём *степенью индивидуализации* воздействия регуляризатора R_i на темы. При $\gamma_i = 0$ абсолютные воздействия не различаются по темам. При $\gamma_i = 1$ они различаются максимально, причём воздействия на большие темы сильнее. Оптимальное значение γ_i предлагается подбирать экспериментальным путём.

Таким образом, вместо задания индивидуальных коэффициентов регуляризации для всех тем предлагается задавать два параметра — τ_i и γ_i. Оба параметра безразмерные и относительные, со значениями, как правило, в отрезке $[0, 1]$.

Регуляризаторы, зависящие от матрицы Θ. Аналогичным образом запишем формулу М-шага (2.16) для суммы регуляризаторов $\tau_i R_i$:

$$\theta_{td} = \underset{t\in T}{\text{norm}}\left(n_{td} + \sum_{i=1}^{k} \tau_i \theta_{td} \frac{\partial R_i}{\partial \theta_{td}}\right).$$

Определим *воздействие* r_{id} регуляризатора R_i на документ d и его *воздействие* r_i на матрицу Θ при $\tau_i = 1$:

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

Введём в формулу М-шага отношения $\frac{n_d}{r_{id}}$ и $\frac{n}{r_i}$, взвешенные *степенью индивидуализации* γ_i воздействия регуляризатора R_i на документы:

$$\theta_{td} = \underset{t\in T}{\text{norm}}\left(n_{td} + \sum_{i=1}^{k} \tau_i \left(\gamma_i \frac{n_d}{r_{id}} + (1-\gamma_i)\frac{n}{r_i}\right) \theta_{td} \frac{\partial R_i}{\partial \theta_{td}}\right).$$

Теперь коэффициент регуляризации τ_i интерпретируется как *относительное воздействие* регуляризатора R_i на матрицу Θ. При $\gamma_i = 0$ абсолютные воздействия регуляризатора не различаются по документам. При $\gamma_i = 1$ они максимально различны, причём воздействия на длинные документы сильнее.

Относительные коэффициенты регуляризации для модальностей. В случае мультимодальной модели (2.19) запишем формулу М-шага (2.24) для модальностей с коэффициентами регуляризации τ_m, $m \in M$:

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

Определим *вес m-й модальности* как число токенов данной модальности в документе d и во всей коллекции, соответственно, r_{md} и r_m:

$$r_{md} = \sum_{t \in T} \sum_{w \in W^m} n_{dw} p_{tdw} = \sum_{w \in W^m} n_{dw}, \qquad r_m = \sum_{d \in D} r_{md}.$$

Будем считать, что первая модальность является основной, и для неё $\tau_1 = 1$. Введём в формулу М-шага отношения весов первой и m-й модальностей $\frac{r_{1d}}{r_{md}}$ и $\frac{r_1}{r_m}$, взвешенные *степенью индивидуализации* γ_m:

$$n_{td} = \sum_{m \in M} \tau_m \left(\gamma_m \frac{r_{1d}}{r_{md}} + (1 - \gamma_m)\frac{r_1}{r_m}\right) \sum_{w \in W^m} n_{dw} p_{tdw}.$$

Теперь коэффициент регуляризации τ_m интерпретируется как *относительное воздействие* модальности m на матрицу Θ. При $\gamma_m = 0$ абсолютные воздействия модальности m не различаются по документам. При $\gamma_m = 1$ они максимально различаются по документам и пропорциональны отношениям числа токенов первой и m-й модальностей в каждом документе. Если $r_{md} = 0$, то данной модальности нет в документе, и слагаемое с $r_{md} = 0$ в знаменателе не возникает.

Разделение суммарного воздействия на сглаживающее и разреживающее. Для большинства регуляризаторов знак производной всегда одинаков. Такие регуляризаторы являются либо всегда сглаживающими, либо всегда разреживающими. Однако для некоторых регуляризаторов тип воздействия на модель может быть не очевиден. В таком случае положительные сглаживающие воздействия r_*^+ и отрицательные (разреживающие) воздействия r_*^- могут оцениваться по отдельности:

$$r_{it}^+ = \sum_{w \in W} \max\left\{\varphi_{wt}\frac{\partial R_i}{\partial \varphi_{wt}}, 0\right\}, \qquad r_i^+ = \sum_{t \in T} r_{it}^+,$$

$$r_{it}^- = \sum_{w \in W} \max\left\{\min\left\{-\varphi_{wt}\frac{\partial R_i}{\partial \varphi_{wt}}, n_{wt}\right\}, 0\right\}, \qquad r_i^- = \sum_{t \in T} r_{it}^-,$$

$$r_{id}^+ = \sum_{t \in T} \max\left\{\theta_{td}\frac{\partial R_i}{\partial \theta_{td}}, 0\right\}, \qquad r_i^+ = \sum_{d \in D} r_{id}^+,$$

$$r_{id}^- = \sum_{t \in T} \max\left\{\min\left\{-\theta_{td}\frac{\partial R_i}{\partial \theta_{td}}, n_{td}\right\}, 0\right\}, \qquad r_i^- = \sum_{d \in D} r_{id}^-.$$

§6.2 Адаптивная траектория регуляризации

TODO: это пока мечта.

7 Мультиграммные тематические модели

TODO: Сергей Царьков, Сергей Стенин.

8 Лингвистическая регуляризация

TODO: Анна Потапенко.

9 Иерархические тематические модели

TODO: Надежда Чиркова.

Список литературы

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

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

[3] *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.

[4] *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.

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

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

[7] *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.

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

[9] *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.

[10] *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.

[11] *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.

[12] *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.

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

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

[15] *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.

[16] *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.

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

[18] *Bolelli L., Ertekin S., Giles C. L.* Topic and trend detection in text collections using latent dirichlet allocation // ECIR. — Vol. 5478 of *Lecture Notes in Computer Science.* — Springer, 2009. — Pp. 776–780.

[19] *Chemudugunta C., Smyth P., Steyvers M.* Modeling general and specific aspects of documents with a probabilistic topic model // Advances in Neural Information Processing Systems. — MIT Press, 2007. — Vol. 19. — Pp. 241–248.

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

[21] *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.

[22] *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.

[23] *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.

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

[25] *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.

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

[27] *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.

[28] *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.

[29] *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 & Knowledge Management. — CIKM '13. — New York, NY, USA: ACM, 2013. — Pp. 2333–2338.

[30] *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.

[31] *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.

[32] *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.

[33] *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.

[34] *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.

[35] *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.

[36] *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.

[37] *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.

[38] *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.

[39] *Newman D., Asuncion A., Smyth P., Welling M.* Distributed algorithms for topic models // *Journal of Machinr Learning Research.* — 2009. — Vol. 10. — Pp. 1801–1828.

[40] *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.

[41] *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.

[42] *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.

[43] *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.

[44] *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.

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

[46] *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.

[47] *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.

[48] *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.

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

[50] *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.

[51] *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.

[52] *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.

[53] *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.

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

[55] *Smola A., Narayanamurthy S.* An architecture for parallel topic models // *Proc. VLDB Endow.* — 2010. — Vol. 3, no. 1-2. — Pp. 703–710.

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

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

[58] *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.

[59] 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.

[60] *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.

[61] *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.

[62] *Vorontsov K. V., Potapenko A. A.* Additive regularization of topic models // *Machine Learning, Special Issue on Data Analysis and Intelligent Optimization.* — 2014.

[63] *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.

[64] *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. — P. 193–202.

[65] *Řehůřek R., Sojka P.* Software framework for topic modelling with large corpora // Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. — Valletta, Malta: ELRA, 2010. — Pp. 45–50.

[66] *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.

[67] *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.

[68] *Wallach H., Mimno D., McCallum A.* Rethinking LDA: Why priors matter // Advances in Neural Information Processing Systems 22: 23rd Annual Conference on Neural Information Processing Systems, Vancouver, BC, Canada / Ed. by Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, A. Culotta. — 2009. — Pp. 1973–1981.

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

[70] *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.

[71] *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.

[72] *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.

[73] *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.

[74] *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.

[75] *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.

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

[77] *Zhai K., Boyd-Graber J., Asadi N., Alkhouja M.* Mr.lda: A flexible large scale topic modeling package using variational inference in mapreduce // Proceedings of the 21st international conference on World Wide Web. — 2012. — Pp. 879–888.

[78] *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.

[79] *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 & Knowledge Management. — CIKM '13. — New York, NY, USA: ACM, 2013. — Pp. 1649–1654.

[80] *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.