Модификации EM-алгоритма для вероятностного тематического моделирования

К. В. Воронцов, А. А. Потапенко
voron@forecsys.ru, anya_potapenko@mail.ru
МФТИ, ВМК МГУ

Вероятностная тематическая модель (ВТМ) строит интерпретируемое представление коллекции текстовых документов, описывая каждый документ дискретным распределением на множестве тем, каждую тему — дискретным распределением на множестве терминов. Рассмотрен обобщенный EM-алгоритм с эвристиками сглаживания, сэмплирования, робастности и разреживания, позволяющий при различных сочетаниях этих эвристик получать как известные тематические модели PLSA (propabilistic latent semantic analysis), LDA (latent Dirichlet allocation), SWB (special words with background), так и новые. Предлагается упрощенный робастный алгоритм, который не требует ни дополнительных вычислительных затрат, ни хранения матрицы параметров шума, и хорошо сочетается с эвристикой разреживания. В экспериментах на двух коллекциях научных публикаций, англоязычной и русскоязычной, подбираются оптимальные сочетания стратегий разреживания и других эвристик. Показывается, что робастная модель без сглаживания позволяет разреживать искомые распределения на 99% без ухудшения качества (перплексии) модели.

Ключевые слова: *вероятностная тематическая модель, байесовский вывод, латентное размещение Дирихле, вероятностный латентный семантический анализ, EM-алгоритм.*

EM-like algorithms for probabilistic topic modeling

K. V. Vorontsov, A. A. Potapenko
Moscow Institute of Physics and Technology, Moscow State University

Probabilistic topic models discover a low-dimensional interpretable representation of text corpora by estimating a multinomial distribution over topics for each document and a multinomial distribution over terms for each topic. A unified family of expectation-maximization (EM) like algorithms with smoothing, sampling, sparsing, and robustness heuristics that can be used in any combinations is considered. The known models PLSA (probabilistic latent semantic analysis), LDA (latent Dirichlet allocation), SWB (special words with background), as well as new ones can be considered as special cases of the presented broad family of models. A new simple robust algorithm suitable for sparse models that do not require to estimate and store a big matrix of noise parameters is proposed. The present authors find experimentally optimal combinations of heuristics with sparsing strategies and discover that sparse robust model without Dirichlet smoothing performs very well and gives more than 99% of zeros in multinomial distributions without loss of perplexity.

Keywords: *probabilistic topic model, bayesian inference, latent dirichlet allocation, probabilistic latent semantic analysis, EM-alorithm.*

Введение

Тематическое моделирование (topic modeling) — одно из современных приложений машинного обучения к анализу текстов, активно развивающееся с конца 1990-х гг. *Вероятностная тематическая модель* коллекции текстовых документов определяет каждую тему как дискретное распределение на множестве терминов, каждый документ — как дис-

кретное распределение на множестве тем. Предполагается, что коллекция документов — это последовательность терминов, выбранных случайно и независимо из смеси этих распределений, и ставится задача восстановления компонент смеси по выборке.

Вероятностные тематические модели применяются для информационного поиска, выявления трендов в научных публикациях и новостных потоках, классификации и категоризации документов, изображений, аудио- и видеосигналов. Многочисленные разновидности и приложения BTM описаны в обзоре [8]. Большинство моделей разрабатываются на основе *латентного размещения Дирихле* LDA [5], с использованием математического аппарата графических моделей и байесовского вывода. Это современный активно развивающийся вероятностный инструментарий, находящий применения повсеместно в задачах анализа данных. Однако в тематическом моделировании он порождает две проблемы.

Во-первых, априорное распределение Дирихле не имеет лингвистических обоснований, не является моделью какого-либо языкового явления, и его применение продиктовано исключительно удобством аналитического интегрирования в байесовском выводе.

Во-вторых, одной из важнейших открытых проблем в теории BTM считается совмещение большого числа функциональных требований в одной модели [8]. Однако байесовский подход оказывается слишком сложным для совмещения более 2–3 требований.

Более простая классическая модель *вероятностного латентного семантического анализа* PLSA [12] не связана с какими-либо параметрическими априорными распределениями. Важной задачей представляется поиск обобщений или модификаций PLSA, имеющих адекватные лингвистические обоснования, обеспечивающие качество моделирования не хуже LDA и упрощающих построение многофункциональных моделей.

В предыдущих работах [1, 19] мы предложили обобщенный алгоритм тематического моделирования, совмещающий эвристики сглаживания, сэмплирования, частого обновления параметров, робастности и разреживания. Комбинируя эти эвристики практически в любых сочетаниях, возможно получать как известные модели PLSA, LDA, CVB0, SWB, так и новые. Эвристики робастности и разреживания имеют очевидный лингвистический смысл и позволяют отказаться от распределения Дирихле без потери качества модели. Было показано, что разреживание позволяет обращать в нуль 90%–95% значений в искомых дискретных распределениях, что позволяет эффективнее решать задачи тематического поиска и классификации больших коллекций текстовых документов. Данная работа является продолжением этих исследований и содержит следующие новые результаты.

1. Предлагаются стратегии постепенного разреживания, позволяющие достигать еще большей разреженности 99% без ухудшения качества модели. Исследуются границы применимости разреживания и его сочетаемость с другими эвристиками.

2. В робастном алгоритме предлагается эвристика постепенного увеличения априорных вероятностей шума и фона.

3. Предлагается упрощенный робастный алгоритм, который разделяет слова на тематические и шумовые без вычисления и хранения матрицы шума.

4. Предлагается новая эвристика разреживания условных распределений тем.

5. Сообщаются результаты экспериментов, в которых модель LDA не имеет значимого преимущества перед PLSA, что противоречит известным экспериментам Д. Блэя и др. [5]. Это объясняется тем, что в [5] сравнивались существенно разные реализации алгоритмов обучения этих двух моделей. Мы же сравниваем реализации, различающиеся только формулой байесовского сглаживания в LDA. Таким образом, мы обосновываем возможность отказа от избыточных вероятностных допущений, присущих LDA, и развития многофункциональных BTM на базе более простого математического аппарата PLSA.

Основные понятия и предположения

В данном разделе сведены основные предположения ВТМ. Предлагаемый набор гипотез отличается от общепринятого тем, в него введены гипотеза разреженности и гипотеза шумовых и фоновых слов. Общепринятая гипотеза об априорных распределениях Дирихле, наоборот, исключена из числа основных предположений и рассматривается далее как одна из дополнительных эвристик. Это отражает точку зрения авторов на то, какие предположения более адекватно описывают особенности текстов на естественном языке.

Пусть D — множество (коллекция) текстовых документов, W — множество (словарь) всех употребляемых в них терминов (слов или словосочетаний). Каждый документ $d \in D$ представляет собой последовательность n_d терминов $(w_1, \ldots, w_{n_d})$ из словаря W. Термин может повторяться в документе много раз.

В роли терминов могут выступать как отдельные слова, так и *ключевые фразы*, выделяемые с помощью тезаурусов [2], статистических критериев [3] или методов машинного обучения [18, 27]. Чтобы не различать формы (склонения, спряжения) слов, на стадии предварительной обработки данных производится либо *лемматизация* — приведение всех терминов к нормальной форме, либо *стемминг* — отбрасывание изменяемых частей слов. Стемминг лучше подходит для английского языка, лемматизация — для русского.

Гипотеза о вероятностном пространстве: *с каждым термином w в документе d может быть связана некоторая тема t из конечного множества тем T, которая не известна; коллекция документов образуется множеством троек (d, w, t), выбираемых случайно и независимо из дискретного распределения $p(d, w, t)$ на конечном множестве $D \times W \times T$.*

Документы $d \in D$ и термины $w \in W$ являются наблюдаемыми переменными, тема $t \in T$ является *латентной* (скрытой) переменной.

Гипотеза независимости элементов выборки: *ни порядок документов в коллекции, ни порядок терминов в документах не важны для выявления тематики.*

Предполагается, что тематику документа можно узнать даже после произвольной перестановки терминов, хотя для человека такой текст теряет смысл. Это предположение называют также гипотезой «мешка слов» (bag of words).

Приняв эту гипотезу, можно перейти к более компактному представлению документа как подмножества $d \subset W$, в котором каждому элементу $w \in d$ поставлено в соответствие число n_{dw} вхождений термина w в документ d.

Гипотеза условной независимости: *условные распределения вероятностей терминов в теме $p(w \mid d, t)$ одинаковы для всех документов $d \in D$ и равны $p(w \mid t)$.*

Это предположение допускает три эквивалентных представления:

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

Вероятностная модель порождения данных. Согласно определению условной вероятности, формуле полной вероятности и гипотезе условной независимости

$$p(w \mid d) = \sum_{t \in T} p(t \mid d)\, p(w \mid t), \tag{2}$$

где $p(t \mid d)$ и $p(w \mid t)$ — искомые распределения. Согласно модели порождения данных (2), коллекция D — это выборка наблюдений (d, w), генерируемых Алгоритмом 1. Процесс порождения последовательности слов документа показан на рис. 1.

Алгоритм 1 Порождение коллекции текстов с помощью вероятностной модели.

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

1: **для всех** $d \in D$
2: задать длину n_d документа d;
3: **для всех** $i = 1, \ldots, n_d$
4: выбрать случайную тему t из распределения $p(t\,|\,d)$;
5: выбрать случайный термин w из распределения $p(w\,|\,t)$;
6: добавить в выборку пару (d, w), при этом тема t «забывается»;

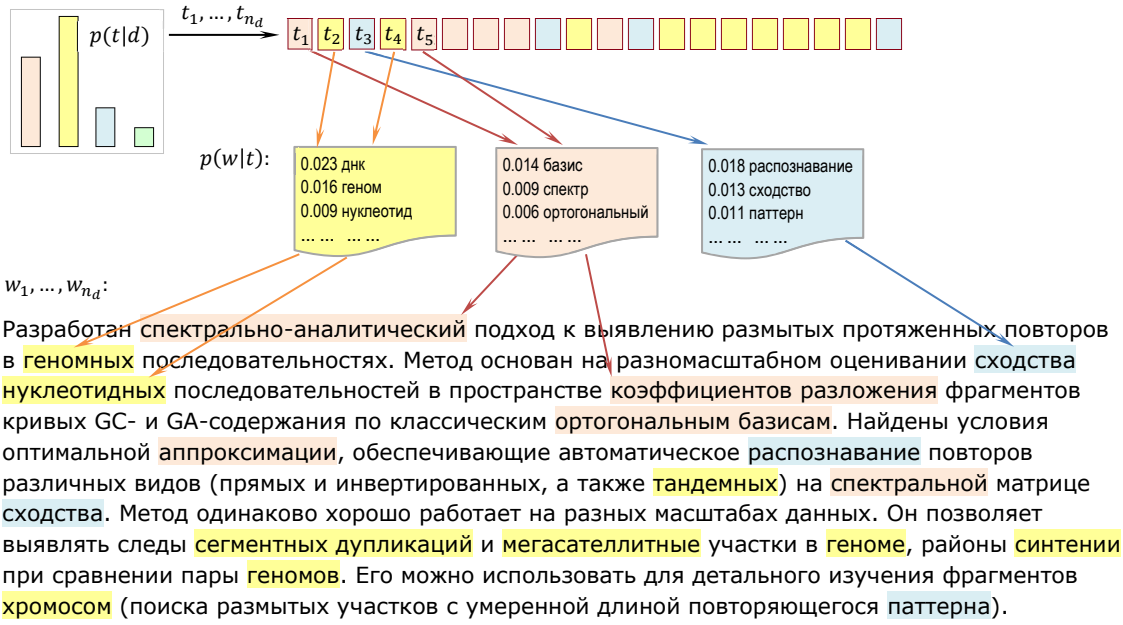


Рис. 1: Процесс порождения текстового документа вероятностной тематической моделью (2)

Построить *тематическую модель* коллекции документов D — значит найти множество тем T, распределения $p(w\,|\,t)$ для всех тем $t \in T$ и $p(t\,|\,d)$ для всех документов $d \in D$.

Распределения $p(t\,|\,d)$ — это сжатые тематические описания документов, которые предполагается использовать для дальнейшего решения задач информационного поиска, классификации, категоризации, аннотирования, суммаризации текстовых документов.

Гипотеза разреженности: *каждый документ d и каждый термин w связан с относительно небольшим числом тем t.*

Каждый документ относится лишь к небольшому числу тем (если же это энциклопедия, то ее лучше разбить на отдельные статьи). Каждая тема состоит из относительно небольшого числа терминов (в работах по филологии почти не встречаются термины из физики, химии, биологии, и многих других наук). Употребление термина в документе, как правило, связано только с одной темой. Таким образом, условные распределения $p(t\,|\,d)$, $p(w\,|\,t)$, $p(t\,|\,d, w)$ должны содержать значительную долю нулевых вероятностей.

Гипотеза шумовых и фоновых слов: *не все слова в тексте тематические.*

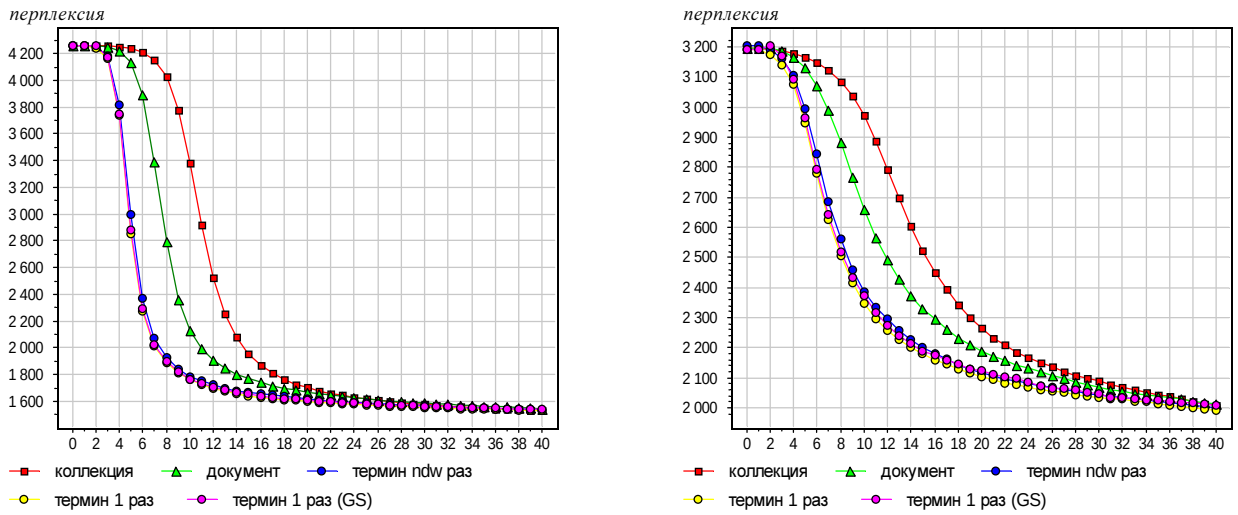


Рис. 2: Зависимость перплексии от числа итераций в стохастическом EM-алгоритме (SEM) при различной частоте обновления параметров φ_{wt}, θ_{td}: после каждого прохода коллекции, после каждого документа, после каждого термина (d, w) по всем n_{dw} его вхождениям, после каждого вхождения термина, GS — с предварительным уменьшением счетчиков как в алгоритме сэмплирования Гиббса. Параметры сглаживания: $\alpha_t = 0{,}5$, $\beta_w = 0{,}01$. Число тем $|T| = 100$

Различаются два вида нетематических слов: *шумовые* слова, специфичные для конкретного документа, и *фоновые* общеупотребительные слова, встречающиеся во многих документах. Фрагмент на рис. 1 содержит много слов, не относящихся ни к одной из тем.

Вообще, *тема* — это статистическое явление, связанное с совместным частым употреблением определенного набора терминов. Термины, употребляемые редко, статистически не значимы, не могут быть тематическими и должны относится к шумовым.

Гипотезы разреженности и шума–фона тесно связаны. Вместе они означают, что в каждом документе возможно отбросить относительно небольшую долю слов (скажем, треть или половину) так, чтобы появление оставшихся терминов описывалось распределениями $p(t \mid d)$, $p(w \mid t)$, разреженными гораздо сильнее (скажем, на 99%). В данной работе приводятся экспериментальные подтверждения этой гипотезы.

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

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

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):

$$\hat{p}(t) = \frac{n_t}{n}, \qquad \hat{p}(w \mid t) = \frac{n_{wt}}{n_t}, \qquad \hat{p}(t \mid d) = \frac{n_{dt}}{n_d}, \qquad \hat{p}(t \mid d, w) = \frac{n_{dwt}}{n_{dw}}, \tag{4}$$

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

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

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

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

В пределе $n \to \infty$ частотные оценки $\hat{p}(\cdot)$, определяемые формулами (3)–(4), стремятся к соответствующим вероятностям $p(\cdot)$.

Метод максимума правдоподобия. Для оценивания параметров тематической модели по коллекции документов D будем максимизировать правдоподобие (плотность распределения) выборки:

$$p(D) = C \prod_{d \in D} \prod_{w \in d} p(d, w)^{n_{dw}} = \prod_{d \in D} \prod_{w \in d} p(w \mid d)^{n_{dw}} \underbrace{C p(d)^{n_{dw}}}_{\text{const}} \to \max,$$

где C — нормировочный множитель, зависящий только от чисел n_{dw}. Отбросим множители C и $p(d)$, не влияющие на положение точки максимума, подставим выражение для $p(w \mid d)$ из (2) и введем для искомых величин обозначения $\theta_{td} = p(t \mid d)$, $\varphi_{wt} = p(w \mid t)$. Прологарифмировав правдоподобие, получим задачу максимизации

$$L(D; \Phi, \Theta) = \sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \varphi_{wt} \theta_{td} \to \max_{\Phi, \Theta} \tag{5}$$

при ограничениях неотрицательности и нормировки:

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

Перплексия (perplexity) является стандартным критерием качества тематических моделей [8]. Это мера несоответствия или «удивленности» модели $p(w \mid d)$ терминам w, наблюдаемым в документах коллекции, определяемая через логарифм правдоподобия (5):

$$\mathscr{P}(D; \Phi, \Theta) = \exp\Big(-\frac{1}{n} L(D; \Phi, \Theta)\Big) = \exp\Big(-\frac{1}{n} \sum_{d \in D} \sum_{w \in d} n_{dw} \ln p(w \mid d)\Big). \tag{6}$$

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

Принято считать, что перплексия, вычисленная по той же коллекции D, по которой строилась модель $\Phi\Theta$, может быть подвержена эффекту переобучения и давать оптимистично заниженные оценки [5]. В наших экспериментах использовалась стандартная методика вычисления *контрольной перплексии* [4]. Коллекция документов изначально разбивалась на две части: обучающую D, по которой строилась модель, и контрольную D',

по которой вычислялась перплексия. После обучения модели векторы φ_t фиксировались, векторы θ_d контрольных документов $d \in D'$ оценивались по первой половине каждого документа, по вторым половинам вычислялась контрольная перплексия. Для разбиения на две половины последовательность терминов $\{w_1, \ldots, w_{n_d}\}$ каждого контрольного документа $d \in D'$ разбивалась после случайной перестановки на две части равной длины. Новые слова, ни разу не встретившиеся в обучающей коллекции D, но попавшие во вторую часть контрольного документа, игнорировались.

Численные эксперименты проводились на двух коллекциях, доступных на викистранице `www.MachineLearning.ru` «Коллекции документов для тематического моделирования». Для обеих коллекций производилась лемматизация и отбрасывались стоп-слова.

Коллекция RuDis содержит $|D| = 2000$ авторефератов диссертаций на русском языке; суммарная длина $n \approx 8.7 \cdot 10^6$, объем словаря $|W| \approx 3 \cdot 10^4$. Контрольная коллекция D' состоит из 200 авторефератов.

Коллекция NIPS содержит $|D| = 1566$ текстов статей научной конференции Neural Information Processing Systems на английском языке; суммарная длина $n \approx 2.3 \cdot 10^6$, объем словаря $|W| \approx 1.3 \cdot 10^4$. Контрольная коллекция D' состоит из 174 документов.

Тематические модели PLSA, LDA, SWB

Данный раздел носит обзорный характер. Вводятся классические тематические модели PLSA [12] и LDA [5]. Предлагается элементарная интерпретация M-шага EM-алгоритма для PLSA. Модель LDA рассматривается как легкое расширение PLSA. Рассматривается робастная модель PLSA-SWB (special words with background) с шумовыми и фоновыми словами. Она устраняет недостатки PLSA и позволяет отказаться от избыточных вероятностных допущений модели LDA [1]. В конце раздела предлагается новая *упрощенная робастная модель*, которая не требует ни дополнительных вычислений, ни памяти для хранения параметров шума и фона.

Вероятностный латентный семантический анализ (probabilistic latent semantic analysis, PLSA) был предложен Томасом Хофманном в [12] и основан на модели (2).

Для решения задачи максимизации правдоподобия (5) в PLSA применяется итерационный процесс, называемый *EM-алгоритмом*, в котором каждая итерация состоит из двух шагов — E (expectation) и M (maximization) [9].

На E-шаге по текущим значениям параметров φ_{wt}, θ_{td} с помощью формулы Байеса вычисляются условные распределения латентных тем $p(t \mid d, w)$ для каждого термина $w \in d$ в каждом документе d:

$$H_{dwt} = p(t \mid d, w) = \frac{p(w \mid t)p(t \mid d)}{p(w \mid d)} = \frac{\varphi_{wt}\theta_{td}}{\sum\limits_{s \in T} \varphi_{ws}\theta_{sd}}. \tag{7}$$

На M-шаге, наоборот, по условным вероятностям тем H_{dwt} вычисляется новое приближение параметров φ_{wt}, θ_{td}. Это легко сделать, если заметить, что величина

$$\hat{n}_{dwt} = n_{dw}p(t \mid d, w) = n_{dw}H_{dwt} \tag{8}$$

оценивает (не обязательно целое) число n_{dwt} вхождений термина w в документ d, связанных с темой t. Просуммировав $\hat{n}_{dwt}$ по документам d и по терминам w, получим оценки

$\hat{n}_{wt}$, $\hat{n}_{dt}$, $\hat{n}_t$, и через них, согласно (4), — частотные оценки условных вероятностей φ_{wt}, θ_{td}:

$$\varphi_{wt} = \frac{\hat{n}_{wt}}{\hat{n}_t}, \qquad \hat{n}_{wt} = \sum_{d \in D} n_{dw} H_{dwt}, \qquad \hat{n}_t = \sum_{w \in W} \hat{n}_{wt}; \tag{9}$$

$$\theta_{td} = \frac{\hat{n}_{dt}}{\hat{n}_d}, \qquad \hat{n}_{dt} = \sum_{w \in d} n_{dw} H_{dwt}, \qquad \hat{n}_d = \sum_{t \in T} \hat{n}_{dt}. \tag{10}$$

Оценки (9)–(10) являются решением задачи максимизации правдоподобия (5) при фиксированных H_{dwt}. Доказательство этого факта можно найти в [13, 1].

Число операций на каждом E- и M-шаге линейно по длине коллекции n и числу тем T.

Линейный проход по коллекции, то есть по всем терминам w во всех документах d, наиболее эффективен, если каждый документ d хранится в виде последовательности пар (w, n_{dw}). Такое представление возможно благодаря гипотезе «мешка слов».

Начальные приближения φ_t и θ_d можно задавать нормированными случайными векторами из равномерного распределения. Другая распространенная рекомендация — пройти по всей коллекции, выбрать для каждой пары (d, w) случайную тему t и вычислить частотные оценки (4) вероятностей φ_{wt} и θ_{td} для всех $d \in D$, $w \in W$, $t \in T$. В экспериментах эти два способа инициализации приводят к схожим результатам.

Недостатком PLSA является невозможность определить, какие из вероятностей θ_{td}, φ_{wt} равны нулю. Согласно формулам (7), (9), (10), если в начальном приближении $\theta_{td} = 0$ (тема t не представлена в документе d) или $\varphi_{wt} = 0$ (термин w не относится к теме t), то нулевое значение будет сохраняться на протяжении всех итераций. Аналогично, исходно ненулевые значения так и остаются ненулевыми. Таким образом, структура разреженности распределений не оптимизируется, а задается через начальное приближение.

Латентное размещение Дирихле. Другим недостатком PLSA принято считать высокую размерность пространства параметров, что может быть причиной переобучения [5]. В задачах машинного обучения для сокращения размерности обычно используется либо *отбор признаков*, приводящий к уменьшению числа параметров, либо *регуляризация* — наложение дополнительных ограничений на параметры. В частности, при *байесовской регуляризации* вводится априорное распределение вероятности в пространстве параметров.

Тематическая модель *латентного размещения Дирихле* (latent Dirichlet allocation, LDA) [5] основана на разложении (2) при дополнительном предположении, что векторы документов $\theta_d = (\theta_{td}) \in \mathbb{R}^{|T|}$ и векторы тем $\varphi_t = (\varphi_{wt}) \in \mathbb{R}^{|W|}$ порождаются распределениями Дирихле с параметрами $\alpha \in \mathbb{R}^{|T|}$ и $\beta \in \mathbb{R}^{|W|}$ соответственно:

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

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

где $\Gamma(z)$ — гамма-функция. Векторы α и β называются *гиперпараметрами*.

Распределение Дирихле принято считать адекватным байесовским регуляризатором в задачах тематического моделирования.

Во-первых, это достаточно широкое параметрическое семейство распределений на единичном симплексе, то есть на множестве дискретных распределений. Если $\alpha_t = 1$ для

всех t, то распределение Дирихле переходит в равномерное. Математическое ожидание и дисперсия t-й координаты вектора θ_d равны, соответственно,

$$\mathsf{E}\theta_{td} = \int \theta_{td} \operatorname{Dir}(\theta_d; \alpha)\, d\theta_d = \frac{\alpha_t}{\alpha_0}, \qquad \mathsf{D}\theta_{td} = \frac{\alpha_t(\alpha_0 - \alpha_t)}{\alpha_0^2(\alpha_0 + 1)}. \tag{11}$$

Векторный параметр α определяет степень разреженности векторов θ_d, порождаемых распределением $\operatorname{Dir}(\theta; \alpha)$. Чем больше α_0, тем сильнее векторы θ_d концентрируется вокруг вектора математического ожидания $\mathsf{E}\theta_d$. Чем меньше α_t, тем сильнее значения θ_{td} концентрируются вокруг нуля. Чем меньше α_0, тем более разрежен вектор θ_d. Поэтому α_t называют *параметрами контраста*.

Во-вторых, модель LDA хорошо подходит для описания кластерных структур. Чем меньше значения гиперпараметров α и β, тем сильнее разрежено распределение Дирихле, и тем дальше отстоят друг от друга порождаемые им векторы. В частности, чем меньше α_0, тем сильнее различаются документы θ_d. Чем меньше β_0, тем сильнее различаются темы φ_t. Векторы $\varphi_t = p(w \mid t)$ в пространстве терминов $\mathbb{R}^{|W|}$ являются центрами тематических кластеров. Элементами кластеров являются эмпирические распределения документов $\hat{p}(w \mid d, t)$. Чем меньше значения гиперпараметров β_w, тем больше межкластерные расстояния по сравнению с внутрикластерными. Таким образом, гиперпараметры позволяют моделировать тематические кластерные структуры различной степени выраженности.

В-третьих, распределение Дирихле является сопряженным к мультиномиальному, что упрощает байесовский вывод апостериорных оценок вероятностей θ_{td} и φ_{wt}.

Недостатком априорного распределения Дирихле является отсутствие убедительных лингвистических обоснований. Предположение, что все векторы θ_d, $d \in D$ порождаются распределением Дирихле, причем одним и тем же, представляется весьма произвольным. То же можно сказать и о порождении векторов распределений φ_t для всех тем $t \in T$.

Второй недостаток заключается в том, что параметры θ_{td}, φ_{wt} и гиперпараметры α_t, β_w не могут обращаться в нуль, что противоречит гипотезе разреженности.

Чтобы получить оценки параметров θ_{td}, φ_{wt} в модели LDA, документ d рассматривается как выборка n_d пар тема–термин $X_d = \{(t_1, w_1), \ldots, (t_{n_d}, w_{n_d})\}$. В каждой паре тема t_i выбирается из дискретного распределения $p(t \mid d) = \theta_{td}$. Следовательно, вероятность встретить каждую из тем t ровно n_{td} раз подчиняется мультиномиальному распределению:

$$p(X_d|\theta_d) = \frac{n_d!}{\prod_t n_{td}!} \prod_t \theta_{td}^{n_{td}}.$$

Распределение Дирихле является *сопряженным* к мультиномиальному. Это означает, что при априорном распределении Дирихле $\theta_d \sim \operatorname{Dir}(\theta; \alpha)$ апостериорное распределение вектора θ_d принадлежит тому же семейству распределений, но с другим значением параметра: $\theta_d|X_d \sim \operatorname{Dir}(\theta; \alpha')$. Действительно, по формуле Байеса

$$p(\theta_d|X_d, \alpha) = \frac{p(X_d|\theta_d) \operatorname{Dir}(\theta_d; \alpha)}{p(X_d)} = C \prod_t \theta_{td}^{n_{td}} \theta_{td}^{\alpha_t - 1} = \operatorname{Dir}(\theta_d; \alpha'), \quad \alpha'_t = \alpha_t + n_{td},$$

где C — нормировочная константа, не зависящая от θ_d.

Оценим случайную величину θ_{td} ее математическим ожиданием (11) по апостериорному распределению:

$$p(t|d, X_d, \alpha) = \int p(t|d) p(\theta_d|X_d, \alpha)\, d\theta_d = \int \theta_{td} \operatorname{Dir}(\theta_d, \alpha')\, d\theta_d = \frac{n_{td} + \alpha_t}{n_d + \alpha_0}. \tag{12}$$

Заменив величину n_{td} ее оценкой $\hat{n}_{td}$, получим сглаженную байесовскую оценку параметра θ_{td} для ЕМ-алгоритма, альтернативную оценке максимума правдоподобия (10):

$$\theta_{td} = \frac{\hat{n}_{dt} + \alpha_t}{\hat{n}_d + \alpha_0}. \tag{13}$$

Аналогично выводится сглаженная байесовская оценка и для φ_{wt}, альтернативная (9):

$$\varphi_{wt} = \frac{\hat{n}_{wt} + \beta_w}{\hat{n}_t + \beta_0}. \tag{14}$$

Частотные оценки условных вероятностей (9)–(10) являются частным случаем сглаженных оценок (14)–(13), что позволяет использовать для обучения моделей LDA и PLSA один и тот же ЕМ-алгоритм. К этому же результату приводят методы сэмплирования Гиббса [20, 25] и вариационной байесовской аппроксимации [21].

Анализ известных алгоритмов обучения LDA показал, что все они являются модификациями ЕМ-алгоритма и отличаются, главным образом, формулой сглаживания частотных оценок вероятностей [4]. Оптимизация гиперпараметров [22, 23] еще сильнее нивелирует различия между алгоритмами. Согласно [11], максимизация апостериорной вероятности в модели LDA при $\alpha = 0$ и $\beta = 0$ приводит к формулам М-шага для модели PLSA.

Далее мы рассмотрим различные модификации ЕМ-алгоритма для обобщенной модели PLSA/LDA и покажем в экспериментах, что использование априорных распределений Дирихле не столь необходимо, как это принято считать. Эвристики разреживания и робастности, а также некоторые особенности реализации ЕМ-алгоритма могут гораздо сильнее влиять на вычислительную эффективность и качество тематической модели.

Робастная тематическая модель формализует предположение, что лишь некоторые слова в текстах относятся к каким-либо темам. Она представляет собой вероятностную смесь трех компонент — тематической, шумовой и фоновой [1]:

$$p(w \,|\, d) = \frac{Z_{dw} + \gamma\pi_{dw} + \varepsilon\pi_w}{1 + \gamma + \varepsilon}; \qquad Z_{dw} = \sum_{t \in T} \varphi_{wt}\theta_{td}. \tag{15}$$

Шумовая компонента $\pi_{dw} \equiv p_{\text{ш}}(w \,|\, d)$ — это слова, специфичные для конкретного документа d, либо редкие термины, относящиеся к темам, слабо представленным в данной коллекции. Отнесение шумовых слов к темам загрязняет распределения $\varphi_{wt} = p(w \,|\, t)$, увеличивает перплексию и искажает тематическую модель.

Фоновая компонента $\pi_w \equiv p_{\text{ф}}(w)$ — это общеупотребительные слова, в частности, стоп-слова, не отброшенные на стадии предварительной обработки. Фоновые слова имеют значимые вероятности во многих темах и только мешают различать темы.

Тематическая компонента Z_{dw} совпадает с моделью PLSA. Если она плохо объясняет избыточную частоту слова в документе, то слово относится к шуму или фону. Параметры γ и ε, ограничивающие долю таких слов, связаны с априорными вероятностями тематической, шумовой и фоновой компонент, равными $1.(1 + \gamma + \varepsilon)$, $\gamma.(1 + \gamma + \varepsilon)$, $\varepsilon.(1 + \gamma + \varepsilon)$ соответственно.

Похожая модель SWB на основе LDA предлагалась в [7]. Основное отличие нашей робастной модели от [7] в том, что она может сочетаться как с LDA, так и с PLSA, и не обязательно связана с сэмплированием Гиббса. Кроме того, мы не вводим априорных распределений Дирихле для параметров π_{dw}, π_w и (γ, ε), полагая, что фоновую и шумовую

компоненты гораздо логичнее разреживать, а не сглаживать. Параметры γ, ε логичнее фиксировать или определять по внешнему критерию качества, так как правдоподобие монотонно возрастает по γ и убывает по ε.

Задача максимизации правдоподобия (5) для модели (15) решена в [1]. По аналогии со стандартным ЕМ-алгоритмом, на Е-шаге для каждой пары (d, w) вычисляются по формуле Байеса условные вероятности тем $H_{dwt} = p(t \mid d, w)$,

$$H_{dwt} = \frac{\varphi_{wt}\theta_{td}}{Z_{dw} + \gamma\pi_{dw} + \varepsilon\pi_w}, \quad t \in T, \tag{16}$$

а также условные вероятности того, что слово w является шумом H_{dw} и фоном H'_{dw}:

$$H_{dw} = \frac{\gamma\pi_{dw}}{Z_{dw} + \gamma\pi_{dw} + \varepsilon\pi_w}; \qquad H'_{dw} = \frac{\varepsilon\pi_w}{Z_{dw} + \gamma\pi_{dw} + \varepsilon\pi_w}. \tag{17}$$

На М-шаге переменные θ_{td} и φ_{wt} вычисляются по прежним формулам (9) и (10) с единственным отличием, что теперь H_{dwt} вычисляются по новой формуле (16). Переменные π_{dw} и π_w вычисляются как частотные оценки условных вероятностей шума и фона:

$$\pi_{dw} = \frac{\nu_{dw}}{\nu_d}, \qquad \nu_{dw} = n_{dw}H_{dw}, \qquad \nu_d = \sum_{w \in d} \nu_{dw},$$
$$\pi_w = \frac{\nu'_w}{\nu'}, \qquad \nu'_w = \sum_{d \in D} n_{dw}H'_{dw}, \qquad \nu' = \sum_{w \in W} \nu'_w,$$

где ν_d и ν' — оценки числа шумовых слов в документе d и фоновых слов во всей коллекции. Эти формулы для π_{dw} и π_w называются *мультипликативным М-шагом*. Они порождают ту же проблему разреженности, что и переменные φ_{wt} и θ_{td}: если в начальном приближении значение π_{dw} или π_w не равно нулю, то оно так и останется ненулевым.

Формула *аддитивного М-шага*, полученная в [1] из условий Куна–Таккера задачи (5), приводит к автоматическому выбору структуры разреженности матрицы $(\pi_{dw})_{D \times W}$:

$$\pi_{dw} = \left(\frac{n_{dw}}{\nu_d} - \frac{Z_{dw} + \varepsilon\pi_w}{\gamma} \right)_+ . \tag{18}$$

Эта формула имеет прозрачную интерпретацию: если термин w в документе d встречается существенно чаще, чем предсказывают тематическая и фоновая компоненты модели, то его появление объясняется особенностями данного документа, и тогда $\pi_{dw} > 0$.

Упрощенная робастная модель. Недостатком предыдущей модели является необходимость подбирать параметры γ, ε и хранить параметры π_{dw}, число которых сопоставимо с размером коллекции. В качестве альтернативы рассмотрим упрощенную робастную модель, в которой фоновая компонента отсутствует, а шумовая компонента π_{dw} включается только когда $Z_{dw} = 0$, то есть когда термин w в документе d не является тематическим:

$$p(w \mid d) = \nu_d Z_{dw} + \big[Z_{dw} = 0\big]\pi_{dw}, \tag{19}$$

где параметр ν_d определяется из условия нормировки $\sum_{w \in W} p(w \mid d) = 1$.

Максимизация правдоподобия (5) снова приводит к частотным оценкам условных вероятностей (9)–(10), но теперь H_{dwt} и $\hat{n}_{dwt}$ оцениваются только по тематическим терминам:

$$\hat{n}_{dwt} = \big[Z_{dw} > 0\big] n_{dw} H_{dwt}.$$

Оптимальное значение π_{dw} достаточно определять только для тех (d,w), при которых $Z_{dw}=0$. Оно также выражается аналитически, $\pi_{dw}=n_{dw}/n_d$, что совпадает с частотной оценкой условной вероятности $p(w\,|\,d)$.

Нормировочный множитель ν_d равен доле тематических терминов в документе:

$$\nu_d=\sum_{w\in W}\big[Z_{dw}>0\big]\pi_{dw}=\frac{1}{n_d}\sum_{w\in d}\big[Z_{dw}>0\big]n_{dw}.$$

Заметим, что параметры π_{dw} и ν_d не нужны для вычисления тематической компоненты модели — матриц Φ и Θ, но могут понадобиться при вычислении перплексии (6), которая непосредственно зависит от $p(w\,|\,d)$.

Упрощенная робастная модель не требует дополнительных затрат памяти или времени. Поэтому в наших экспериментах она используется всегда, когда возможно обнуление тематической компоненты Z_{dw}, если явно не указано, что используется робастная модель (15).

Обобщенный EM-алгоритм и его модификации

В данном разделе рассматриваются рациональный, обобщенный и стохастический варианты EM-алгоритма [1]. Алгоритм сэмплирования Гиббса рассматривается как частный случай стохастического EM-алгоритма, применимый не только к модели LDA, но также и к PLSA. Тем самым мы показываем, что PLSA и LDA отличаются только эвристикой сглаживания, а многочисленные варианты EM-алгоритма одинаково применимы к обеим моделям. Предлагается несколько новых эвристик: принудительное разреживание условных распределений тем $p(t\,|\,d,w)$, постепенное увеличение априорных вероятностей шума и фона в робастной модели, простая и сложная стратегии разреживания матриц Φ и Θ. По сравнению с [1] существенно расширен состав экспериментов по подбору оптимального сочетания эвристик в EM-алгоритме.

Рациональный EM-алгоритм. Начнем с простой реорганизации шагов EM-алгоритма, чтобы избежать хранения трехмерной матрицы H_{dwt}. Заметим, что переменные $\hat{n}_{wt}$, $\hat{n}_{dt}$, $\hat{n}_t$ вычисляются на M-шаге в цикле по всем документам $d\in D$ и всем терминам $w\in d$. Внутри этого цикла переменные H_{dwt} будем вычислять непосредственно в тот момент, когда они понадобятся. Переменную $\hat{n}_d$ можно не вычислять, поскольку $\hat{n}_d=n_d$. Тогда E-шаг встраивается внутрь M-шага без дополнительных вычислительных затрат. Этот вариант EM-алгоритма будем называть *рациональным*, он показан в Алгоритме 2.

Условия остановки в данном алгоритме и всех последующих не сформулированы. В наших экспериментах 40 итераций всегда было достаточно для сходимости перплексии на обучающей и контрольной выборках.

Обобщенный EM-алгоритм. Известно, что на каждом M-шаге нет необходимости слишком точно решать задачу максимизации правдоподобия. Достаточно сместиться в направлении максимума и затем выполнить E-шаг. Этот вариант EM-алгоритма называется *обобщенным EM-алгоритмом* (generalized EM-algorithm, GEM) [9]. Другое обобщение состоит в том, что E-шаг выполняется только для части скрытых переменных, после этого M-шаг выполняется только для тех переменных, значения которых зависят от изменившихся скрытых переменных [17]. Для обобщенных EM-алгоритмов справедливы те же обоснования сходимости, что и для стандартного EM-алгоритма.

Алгоритм 2 PLSA-EM: рациональный EM-алгоритм для модели PLSA.

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

пока не выполнится критерий остановки, повторять итерации:
  обнулить $\hat{n}_{wt}$, $\hat{n}_{dt}$, $\hat{n}_t$ для всех $d \in D,\ w \in W,\ t \in T$;
  для всех $d \in D,\ w \in d$
    $Z := \sum_{t \in T} \varphi_{wt}\theta_{td}$;
    для всех $t \in T$ таких, что $\varphi_{wt}\theta_{td} > 0$
      увеличить $\hat{n}_{wt}$, $\hat{n}_{dt}$, $\hat{n}_t$ на $\delta = n_{dw}\varphi_{wt}\theta_{td}/Z$;
  $\varphi_{wt} := \hat{n}_{wt}/\hat{n}_t$ для всех $w \in W,\ t \in T$;
  $\theta_{td} := \hat{n}_{dt}/n_d$ для всех $d \in D,\ t \in T$;

Алгоритм 3 PLSA-GEM: обобщенный EM-алгоритм для модели PLSA.

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

обнулить $\hat{n}_{wt}$, $\hat{n}_{dt}$, $\hat{n}_t$, $\hat{n}_d$, n_{dwt} для всех $d \in D,\ w \in W,\ t \in T$;
пока не выполнится критерий остановки, повторять итерации:
  для всех $d \in D,\ w \in d$
    $Z := \sum_{t \in T} \varphi_{wt}\theta_{td}$;
    для всех $t \in T$ таких, что $n_{dwt} > 0$ или $\varphi_{wt}\theta_{td} > 0$
      $\delta := n_{dw}\varphi_{wt}\theta_{td}/Z$;
      увеличить $\hat{n}_{wt}$, $\hat{n}_{dt}$, $\hat{n}_t$, $\hat{n}_d$ на $(\delta - n_{dwt})$;
      $n_{dwt} := \delta$;
    если пора обновить параметры Φ, Θ **то**
      $\varphi_{wt} := \hat{n}_{wt}/\hat{n}_t$ для всех $w \in W,\ t \in T$ таких, что $\hat{n}_{wt}$ изменился;
      $\theta_{td} := \hat{n}_{dt}/\hat{n}_d$ для всех $d \in D,\ t \in T$ таких, что $\hat{n}_{dt}$ изменился;

Обобщенный EM-алгоритм для PLSA/LDA отличается от стандартного более частым обновлением параметров θ_{td} и φ_{wt} по текущим значениям счетчиков $\hat{n}_{wt}$ и $\hat{n}_{dt}$. Возможные варианты обновлений — после каждого документа или заданного числа документов, после каждого термина (d, w) или заданного числа терминов, после каждого вхождения термина.

В Алгоритме 2 обновления происходят после каждого прохода коллекции.

В Алгоритме 3 моменты обновления выбираются на шаге 9. В экспериментах на достаточно больших коллекциях частые обновления ускоряют сходимость, рис. 2.

При обновлении после каждого термина или каждого вхождения можно не хранить значения φ_{wt}, θ_{td}, а вычислять их каждый раз как частное двух счетчиков.

На первой итерации (т. е. при первом проходе коллекции) частые обновления не делаются, чтобы в счетчиках накопилась информация по всей коллекции. В противном случае оценки параметров θ_{td} и φ_{wt} по начальному фрагменту выборки могут оказаться хуже начального приближения. Начиная со второй итерации, для каждого термина (d, w) из счетчиков $\hat{n}_{wt}$ и $\hat{n}_{dt}$ вычитается n_{dwt} — то самое значение δ, которое было к ним прибавлено при обработке термина (d, w) на предыдущей итерации. Таким образом, счетчики $\hat{n}_{wt}$ и $\hat{n}_{dt}$ всегда содержат результат последнего однократного прохода всей коллекции.

Недостатком Алгоритма 3 является необходимость хранить массив значений n_{dwt}, $t \in T$ для каждого термина (d, w). Расход памяти объема $O(n|T|)$ может оказаться неприемлемым даже при небольшом числе тем. С другой стороны, согласно гипотезе разреженности, этот массив должен состоять преимущественно из нулей. Далее рассматриваются несколько альтернативных способов разреживания распределений $p(t \mid d, w)$.

Принудительное разреживание условных распределений тем. Стратегия *максимального разреживания* распределений $p(t \mid d, w)$ на первый взгляд представляется наиболее естественной: для каждого термина (d, w) игнорируются темы с наименьшими вероятностями, остаются только s тем с наибольшими значениями n_{dwt}.

Эксперименты показывают, что эта стратегия приводит к накоплению систематической ошибки и расходимости (рис. 3). На первых же итерациях возникает сильная (свыше 90%) разреженность распределений $\varphi_{wt} = p(w \mid t)$, которые к этому моменту еще не сошлись. Значения φ_{wt}, оказавшиеся равными нулю, далее так и остаются нулевыми. Эвристики сглаживания или включения разреживания с 10-й итерации не решают проблему.

Более удачной оказалась стратегия *постепенного разреживания*, когда в каждом распределении $p(t \mid d, w)$ обнуляется заданная доля r наименьших ненулевых значений и производится перенормировка. Эксперименты показали, что при $r \leqslant 0{,}2$ и включении разреживаний начиная с 10-й итерации расходимость не возникает и финальная перплексия мало отличается от случая $r = 0$. При этом постепенно увеличивается разреженность распределений θ_{td} (до 0,5 и выше) и распределений φ_{wt} (немного ниже 0,5).

Робастные алгоритмы более устойчивы к постепенному разреживанию распределений $p(t \mid d, w)$, для них параметр разреживания можно увеличивать до $r = 0{,}5$, при этом разреженность φ_{wt} достигает почти 0,9, разреженность θ_{td} достигает 0,7.

Стохастический ЕМ-алгоритм (stochastic EM-algorithm, SEM) [6] приводит к другой адекватной стратегии разреживания распределений $p(t \mid d, w)$. Для каждой пары (d, w) распределение $p(t \mid d, w)$ используется только для сэмплирования s случайных тем t_{dwi}, $i = 1, \ldots, s$, после чего оно «забывается». В формулах М-шага вместо распределения $p(t \mid d, w)$ используется его несмещенная эмпирическая оценка:

$$\hat{p}(t \mid d, w) = \frac{1}{s} \sum_{i=1}^{s} \big[t_{dwi} = t\big]. \tag{20}$$

Тем самым обеспечивается несмещенность оценок φ_{wt}, θ_{td} и сходимость ЕМ-алгоритма. Объем s сэмплируемых выборок является параметром метода.

Модификация Алгоритма 3, трансформирующая его в стохастический обобщенный ЕМ-алгоритм (PLSA-SGEM), состоит из трех изменений:

1. перед шагом 5 сэмплируется s тем t_{dwi}, $i = 1, \ldots, s$ из $p(t \mid d, w)$;
2. на шаге 5 цикл по всем $t \in T$ заменяется циклом по $t = t_{dwi}$, $i = 1, \ldots, s$;
3. на шаге 6 вычисляется $\delta := n_{dw}/s$.

Эксперименты показывают, что достаточно сэмплировать совсем небольшое число тем, около 5 тем обычно достаточно (табл. 1 и 2). Эта эвристика, названная *экономным сэмплированием* [1], сокращает затраты времени и памяти в тех случаях, когда средняя по коллекции величина n_{dw} превышает s.

В эксперименте проверялась также гипотеза, что число тем, связанных с парой (d, w), не должно превышать числа употреблений данного слова n_{dw}. Для этого производилось

сэмплирование $\min\{s, n_{dw}\}$ тем, однако результаты для этой эвристики немного хуже, чем при сэмплировании ровно s тем.

Робастная модель менее чувствительна к выбору параметра s (табл. 2).

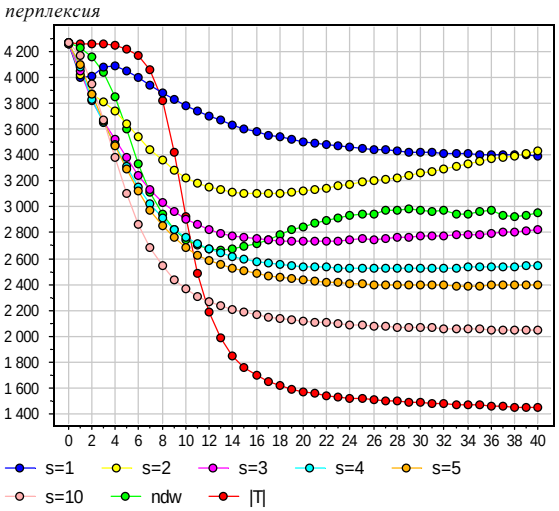
Качественные выводы, которые можно сделать по обучающей и по контрольной выборке, совпадают (табл. 1 и 2). В дальнейших экспериментах это тоже всегда так, но данные по обучающей выборке не показываются в таблицах и графиках.

Сэмплирование Гиббса. При $s = n_{dw}$ стохастический EM-алгоритм со сглаживанием (LDA-SEM) становится похож на *сэмплирование Гиббса* (Gibbs Sampling, GS) — один из основных методов обучения вероятностных тематических моделей [20, 25], см. Алгоритм 4. Алгоритмы LDA-SEM и LDA-GS отличаются несколькими деталями, которые, как показывают эксперименты, почти не влияют на качество модели.

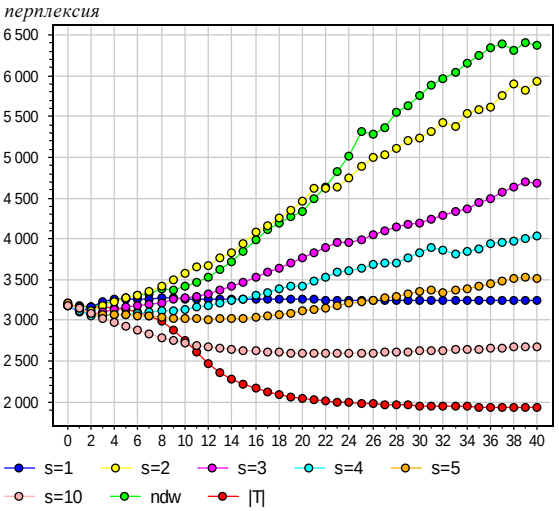
1. В LDA-GS число сэмплирований тем $s = n_{dw}$ для каждой пары (d, w). Согласно описанным выше экспериментам, можно также сэмплировать фиксированное число тем.

2. В LDA-GS параметры φ_{wt} и θ_{td} обновляются предельно часто — после обработки каждого вхождения термина w в документ d. В LDA-SEM обновления могут производиться с любой частотой. Эксперименты показывают, что частота обновления влияет только на скорость сходимости, но почти не влияет на значение контрольной перплексии в конце итераций, рис. 2. По результатам эксперимента можно рекомендовать обновления после каждого термина или после каждого вхождения термина, как в LDA-GS.

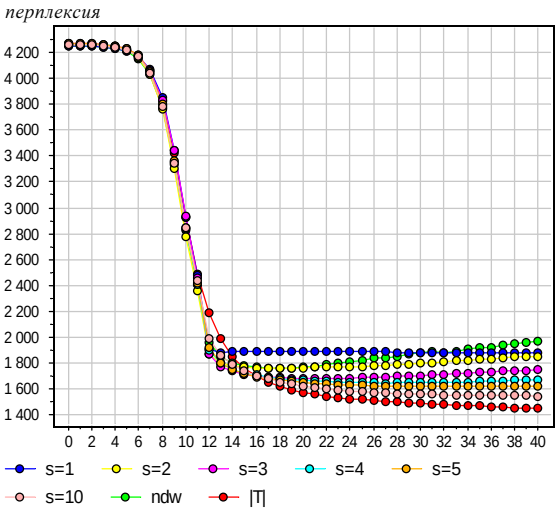
3. В LDA-GS перед сэмплированием счетчики уменьшаются на единицу (шаг 5). Тем самым при оценивании распределений не учитывается i-е вхождение термина w в документ d, для которого сэмплируется тема t_{dwi}. Из теории следует, что эта особенность исключительно важна [25]. Однако в экспериментах с коллекциями достаточно больших размеров оказывается, что она не влияет на качество модели — кривые «термин 1 раз» и «термин 1 раз (GS)» на рис. 2 практически совпадают. Можно одновременно уменьшать счетчики для старой темы и увеличивать для новой, как в Алгоритме 3.



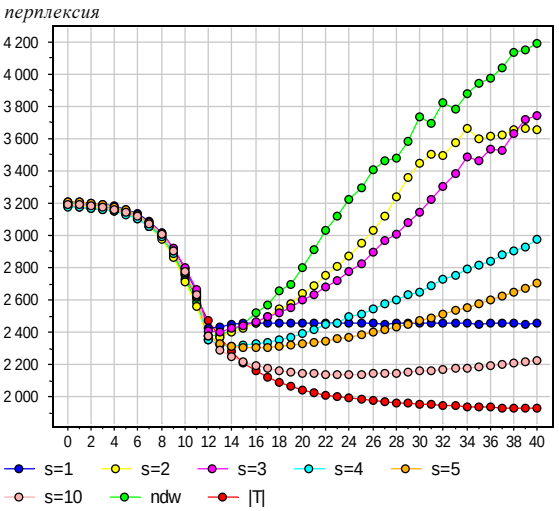
RuDis, разреживание с 1-й итерации



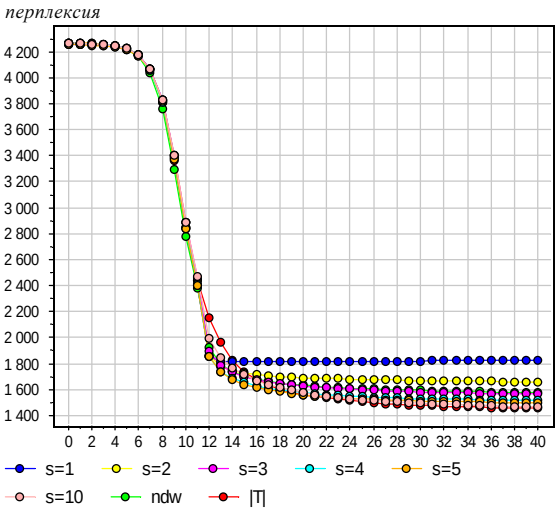
NIPS, разреживание с 1-й итерации



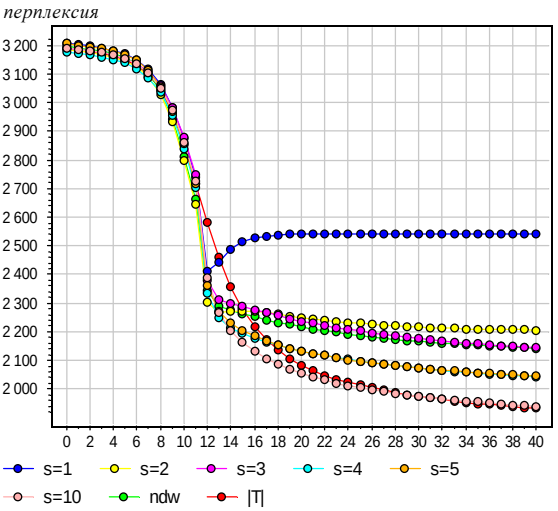
RuDis, разреживание с 10-й итерации



NIPS, разреживание с 10-й итерации



RuDis, с 10-й итерации, сглаживание LDA



NIPS, с 10-й итерации, сглаживание LDA

Рис. 3: Зависимость перплексии от числа итераций в рациональном EM-алгоритме при максимальном разреживании $p(t \mid d, w)$. Параметр разреживания: $s = 1, 2, 3, 4, 5, 10, n_{dw}$, при $s = |T|$ разреживания нет. Параметры сглаживания: $\alpha_t = 0.5$, $\beta_w = 0.01$. Число тем $|T| = 100$.

Таблица 1: Стохастический EM-алгоритм для модели LDA. Зависимость перплексии на обучении и контроле от объема s сэмплированной выборки (40 итераций, $\alpha_t = 0{,}5$, $\beta_w = 0{,}01$)

RuDis: s фиксирован

s	обуч.	конт.
n_{dw}	1367	1535
1	1707	1874
2	1547	1705
3	1463	1628
4	1407	1552
5	1383	1559
10	1295	1480

RuDis: $\min\{s, n_{dw}\}$

s	обуч.	конт.
n_{dw}	1367	1535
1	1724	1894
2	1575	1730
3	1507	1673
4	1479	1647
5	1459	1603
10	1418	1571

NIPS: s фиксирован

s	обуч.	конт.
n_{dw}	1506	2002
1	1796	2326
2	1616	2120
3	1513	2006
4	1473	1981
5	1430	1946
10	1326	1874

NIPS: $\min\{s, n_{dw}\}$

s	обуч.	конт.
n_{dw}	1506	2002
1	1791	2313
2	1647	2157
3	1591	2101
4	1562	2052
5	1547	2052
10	1517	2019

Таблица 2: Стохастический EM-алгоритм для робастной модели LDA. Зависимость перплексии на обучении и контроле от объема s сэмплированной выборки (40 итераций, $\alpha_t = 0{,}5$, $\beta_w = 0{,}01$)

RuDis: s фиксирован

s	обуч.	конт.
n_{dw}	717	794
1	777	857
2	754	830
3	736	815
4	728	804
5	724	799
10	715	789

RuDis: $\min\{s, n_{dw}\}$

s	обуч.	конт.
n_{dw}	717	794
1	773	850
2	748	821
3	737	811
4	731	807
5	728	801
10	722	800

NIPS: s фиксирован

s	обуч.	конт.
n_{dw}	1110	1363
1	1270	1544
2	1171	1442
3	1140	1414
4	1103	1375
5	1087	1352
10	1053	1317

NIPS: $\min\{s, n_{dw}\}$

s	обуч.	конт.
n_{dw}	1110	1363
1	1263	1530
2	1185	1464
3	1167	1441
4	1150	1423
5	1133	1398
10	1121	1393

Алгоритм 4 LDA-GS: сэмплирование Гиббса для тематической модели LDA.

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

обнулить $\hat{n}_{wt}$, $\hat{n}_{dt}$, $\hat{n}_t$ для всех $d \in D$, $w \in W$, $t \in T$;
пока не выполнится критерий остановки, повторять итерации:
  для всех $d \in D$, $w \in d$, $i = 1, \ldots, n_{dw}$
    если не первая итерация **то**
      $t := t_{dwi}$; уменьшить $\hat{n}_{wt}$, $\hat{n}_{dt}$, $\hat{n}_t$ на 1;
    сэмплировать тему t_{dwi} из $p(t \mid d, w) \propto (\hat{n}_{dt} + \alpha_t)(\hat{n}_{wt} + \beta_w)/(\hat{n}_t + \beta_0)$;
    $t := t_{dwi}$; увеличить $\hat{n}_{wt}$, $\hat{n}_{dt}$, $\hat{n}_t$ на 1;
$\varphi_{wt} = (\hat{n}_{wt} + \beta_w)/(\hat{n}_t + \beta_0)$ для всех $t \in T$, $w \in W$;
$\theta_{td} := (\hat{n}_{dt} + \alpha_t)/(n_d + \alpha_0)$ для всех $d \in D$, $t \in T$;

Таким образом, главной особенностью алгоритма сэмплирования Гиббса, как и стохастического EM-алгоритма, является *эвристика сэмплирования* — замена распределения тем $p(t \mid d, w)$ его разреженным эмпирическим аналогом (20). Хотя в литературе алгоритм

Алгоритм 5 PLSA-REM: робастный рациональный EM-алгоритм для модели PLSA.

Вход: коллекция D, число тем $|T|$, начальные приближения Θ, Φ, параметры γ, ε;
Выход: распределения: матрицы (φ_{wt}), (θ_{td}), (π_{dw}) и вектор (π_w);

инициализировать $\pi_{dw} := n_{dw}/n_d$; $\pi_w := n_w/n$; для всех $d \in D$, $w \in W$;
пока не выполнится критерий остановки, повторять итерации:
  обнулить $\hat{n}_{wt}$, $\hat{n}_t$, ν'_w, ν, ν' для всех $w \in W$, $t \in T$;
  для всех $d \in D$
    обнулить $\hat{n}_{dt}$, $\hat{n}_d$, ν_d для всех $w \in W$, $t \in T$;
    для всех $w \in d$
      $Z := \gamma\pi_{dw} + \varepsilon\pi_w + \sum_{t\in T} \varphi_{wt}\theta_{td}$;
      увеличить $\hat{n}_{wt}$, $\hat{n}_{dt}$, $\hat{n}_t$, $\hat{n}_d$ на $n_{dw}\varphi_{wt}\theta_{td}/Z$ для всех $t \in T$;
      увеличить ν_d, ν на $\nu_{dw} := n_{dw}\gamma\pi_{dw}/Z$;
      увеличить ν'_w, ν' на $n_{dw}\varepsilon\pi_w/Z$;
    $\theta_{td} := \hat{n}_{dt}/\hat{n}_d$ для всех $t \in T$;
    $\pi_{dw} := \nu_{dw}/\nu_d$ для всех $w \in d$;
  $\varphi_{wt} := \hat{n}_{wt}/\hat{n}_t$ для всех $w \in W$, $t \in T$;
  $\pi_w := \nu'_w/\nu'$ для всех $w \in W$;

сэмплирования Гиббса принято связывать с моделью LDA, он также в равной степени применим и к модели PLSA.

Эвристика *постепенного разреживания* является альтернативой сэмплированию. Обе эвристики легко встраиваются в любой EM-подобный алгоритм и сочетаются с другими эвристиками. Недостатком постепенного разреживания является необходимость хранения плотных массивов n_{dwt} на начальных итерациях.

В итоге рекомендуется либо рациональный EM-алгоритм с обновлением распределений Φ после каждого прохода коллекции, либо стохастический EM-алгоритм с экономным сэмплированием и обновлением после каждого термина, либо сэмплирование Гиббса.

Робастный EM-алгоритм. В отличие от рационального EM-алгоритма, n_{dw} вхождений термина w в документ d распределяются не только между темами $t \in T$, но также между шумовой и фоновой компонентами пропорционально вероятностям

$$\tilde{H}_{dw} = \big(\tfrac{1}{Z}\varphi_{wt}\theta_{td},\ t \in T;\ \ \tfrac{1}{Z}\gamma\pi_{dw};\ \ \tfrac{1}{Z}\varepsilon\pi_w\big),$$

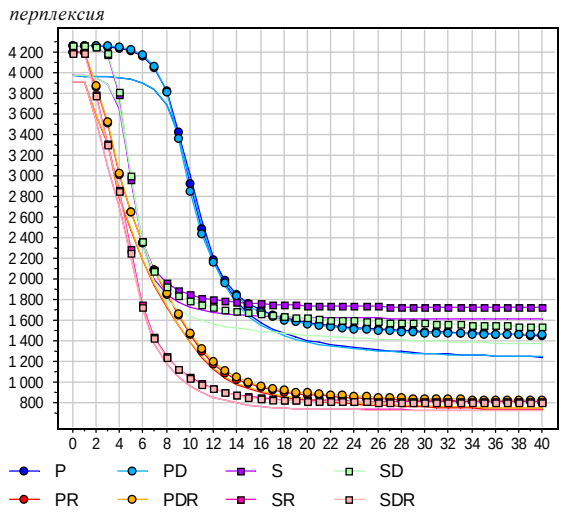
где Z — нормирующий множитель, см. Алгоритм 5.

В экспериментах использовался также стохастический робастный алгоритм с параметром сэмплирования $s = n_{dw}$. Все алгоритмы сравнивались в двух вариантах: с несмещенными оценками (9)–(10) и сглаженными оценками (13)–(14) параметров φ_{wt} и θ_{td}.

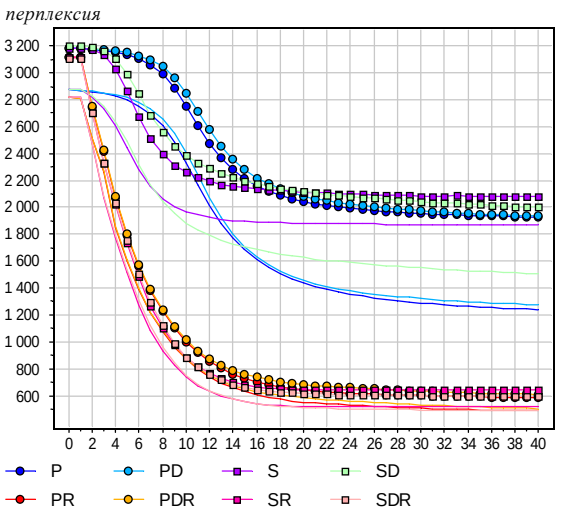
При вычислении перплексии на документах d контрольной выборки D' параметры φ_{wt} и π_w оценивались по обучающей выборке D, параметры θ_{td} и ν_d оценивались по первой половине документа d', параметры π_{dw} оценивались для каждой пары (d, w) согласно (18). Перплексия вычислялась по вторым половинам d'' контрольных документов.

Сравнение восьми алгоритмов, образуемых всеми комбинациями эвристик сглаживания, робастности и сэмплирования (рис. 4) позволяет сделать следующие выводы:

— для обеих задач робастные алгоритмы существенно превосходят неробастные и гораздо меньше переобучаются;



RuDis



NIPS

Рис. 4: Зависимость контрольной перплексии от числа итераций для всевозможных сочетаний эвристик: D — сглаживание Дирихле ($\alpha_t = 0{,}5$, $\beta_w = 0{,}01$); R — робастность ($\gamma = 0{,}3$, $\varepsilon = 0{,}01$); S — сэмплирование ($s = n_{dw}$), P — пропорциональное распределение (8); $|T| = 100$. Тонкие кривые без точек — перплексия обучающей выборки

Таблица 3: Контрольная перплексия $\mathscr{P}$ и оценки апостериорной вероятности шума $\hat{p}_{\text{ш}}$ и фона $\hat{p}_{\text{ф}}$ при различных значениях γ и ε (после 40 итераций, $|T| = 100$)

RuDis, $\varepsilon = 0{,}01$:

γ	$\mathscr{P}$	$\hat{p}_{\text{ш}}$
0	1540	0,000
0,001	1434	0,026
0,01	1277	0,090
0,05	1076	0,196
0,1	974	0,266
0,3	805	0,413
0,5	750	0,498

RuDis, $\gamma = 0{,}3$:

ε	$\mathscr{P}$	$\hat{p}_{\text{ф}}$
0	797	0,000
0,01	794	0,006
0,05	798	0,027
0,1	809	0,049
0,2	823	0,086
0,3	841	0,116
0,5	870	0,165

NIPS, $\varepsilon = 0{,}01$:

γ	$\mathscr{P}$	$\hat{p}_{\text{ш}}$
0	2001	0,000
0,001	1763	0,044
0,01	1381	0,152
0,05	991	0,296
0,1	818	0,377
0,3	604	0,527
0,5	525	0,598

NIPS, $\gamma = 0{,}3$:

ε	$\mathscr{P}$	$\hat{p}_{\text{ф}}$
0	598	0.000
0,01	596	0,005
0,05	605	0,023
0,1	613	0,043
0,2	630	0,079
0,3	640	0,109
0,5	668	0,157

— сэмплирование (20) немного хуже пропорционального распределения (8);
— сэмплирование без сглаживания может приводить к увеличению перплексии.

Величина переобучения (разность перплексии на обучающей и контрольной выборке) больше зависит от задачи, чем от алгоритма. Сравнение алгоритмов по перплексии на обучающей выборке приводит к тем же качественным выводам, что и их сравнение по перплексии на контрольной выборке. По всей видимости, для сравнения алгоритмов не нужна столь сложная методика разделения контрольных документов для вычисления перплексии; вполне достаточно вычислять перплексию только на обучающей выборке.

Возможны два варианта реализации M-шага — мультипликативный и аддитивный. В экспериментах на обеих задачах они не дают значимых различий перплексии.

Возможны два варианта определения роли каждого слова (d, w) при сэмплировании из распределения $\tilde{H}_{dw}$. В первом варианте роли распределяются между компонентами

тем, шума и фона «мягко», пропорционально их вероятностям, затем сэмплируются темы. Во втором варианте сэмплирование производится из всего распределения $\tilde{H}_{dw}$, в результате каждому слову «жестко» приписывается одна из трех взаимоисключающих ролей. В экспериментах эти два варианта также не дают значимых различий перплексии.

Зависимость перплексии от параметров γ и ε, как правило, монотонная, причем параметр γ гораздо сильнее влияет на перплексию, чем ε, см. табл. 3. С ростом γ перплексия уменьшается, так как компонента шума близка к униграммной модели документа, $\pi_{dw} \approx n_{dw}/n_d$, которая наиболее точно предсказывает вероятности слов $p(w \mid d)$, однако не является тематической. С ростом ε перплексия увеличивается, так как компонента фона близка к униграммной модели коллекции, $\pi_w \approx n_w/n$, которая хуже предсказывает вероятности слов $p(w \mid d)$, чем тематическая модель. Оценки апостериорных вероятностей шума $\hat{p}_{\text{ш}} = \nu/n$ и фона $\hat{p}_{\text{ф}} = \nu'/n$ также зависят от γ и ε монотонно. Следовательно, оптимальные значения параметров γ и ε должны определяться по внешним критериям качества той прикладной задачи, для решения которой строится тематическая модель.

На рис. 5 показаны зависимости перплексии и апостериорной вероятности шума от числа итераций при $\gamma = 0{,}0,\ 0{,}001,\ 0{,}3$ и при постепенном увеличении γ от $\gamma_0 = 0{,}001$ до $\gamma_1 = 0{,}3$ на первых $i_1 = 20$ итерациях:

$$\gamma = \gamma_0 + \frac{(\gamma_1 - \gamma_0)i^2}{i^2 + (i_1 - i)^2}.$$

Эвристика постепенного увеличения априорной вероятности шума позволяет достичь немного лучшего значения перплексии. Это можно объяснить тем, что шумовая компонента слишком агрессивно отбирает слова на первых же итерациях, когда тематическая компонента еще не успела сойтись.

Разреживающий ЕМ-алгоритм. Гипотеза разреженности предполагает, что коллекция порождается дискретными распределениями $\varphi_{wt} = p(w \mid t)$ и $\theta_{td} = p(t \mid d)$, в которых подавляющее большинство вероятностей равны нулю. Следствием этого является также и разреженность распределений $H_{dwt} = p(t \mid d, w)$. Обнуление значительной доли вероятностей φ_{wt} и θ_{td} позволяет ускорить ЕМ-алгоритм и хранить тематическую модель в более сжатом виде, открывая возможности для обработки очень больших коллекций.

Модель PLSA не оптимизирует структуру разреженности распределений и требует задавать ее через начальное приближение. Отдельные значения φ_{wt} и θ_{td} могут в ходе итераций сами собой приближаться к нулю, но, как правило, их доля недостаточна для получения выигрыша в производительности.

Модель LDA также не является разреженной — априорные распределения Дирихле запрещают вероятностям φ_{wt} и θ_{td} и гиперпараметрам β_w и α_t принимать нулевые значения. При стремлении гиперпараметров к нулю распределения Дирихле порождают векторы φ_t и θ_d, компоненты которых стремятся к нулю, но никогда не обращаются в нуль. Сглаженные оценки (14) и (13), используемые в LDA, менее разрежены, чем несмещенные частотные оценки (9) и (10), используемые в PLSA.

Известные подходы к разреживанию LDA требуют введения дополнительных параметров и усложнения ЕМ-алгоритма. В [10] предлагается хранить не сами значения φ_{wt} и θ_{td}, а только их разности с фоновыми распределениями. В [24] предполагается, что каждая тема описывается распределением Дирихле на подмножестве слов, заданном бинарными

переменными b_{wt} из распределения Бернулли. Сглаженность и разреженность регулируется независимо параметрами распределения Дирихле и распределения Бернулли. Недостатком данной модели является большое число дополнительных скрытых переменных, которые усложняют обучение. В [14] вводится распределение псевдо-Дирихле, которое строится путем расширения области определения распределения Дирихле и имеет ограниченную плотность, в то время как распределение Дирихле не ограничено в случае $\alpha < 1$, что и приводит к запрету нулевых значений φ_{wt} и θ_{td}.

В данной работе исследуются различные стратегии *принудительного разреживания*, когда в конце каждой итерации (полного прохода всей коллекции D) обнуляется некоторое количество наименьших значений φ_{wt} и θ_{td}. Эвристика разреживания не совместима со сглаживанием и применяется только к PLSA, т. е. при $\beta_w = 0,\ \alpha_t = 0$.

Предварительные эксперименты показали, что одновременное обнуление более 50% элементов является слишком сильным стрессом для модели и может вызывать расходимость ЕМ-алгоритма. Поэтому предлагается разреживать матрицы Φ и Θ постепенно, придерживаясь одной из следующих стратегий.

Простая стратегия: в каждом из распределений φ_t, θ_d обнуляется заданная доля r наименьших *ненулевых* значений. После обнуления производится перенормировка распределений. Число обнуляемых значений сокращается от итерации к итерации, поскольку доля берется от числа ненулевых значений. Обнуления прекращаются, когда в распределении остается $\lfloor r^{-1} \rfloor$ ненулевых значений. Недостатком этой стратегии является стремление к выравниванию доли ненулевых значений во всех распределениях, что представляется довольно странным ограничением.

Сложная стратегия устраняет этот недостаток. В каждом из распределений φ_t, θ_d обнуляется максимальное число наименьших значений, так, чтобы оно не превышало $r|W|$ и $r|T|$ соответственно, и сумма обнуляемых значений не превышала заданного порога R_φ или R_θ для распределений φ_t или θ_d соответственно.

Разреживания включаются, начиная с итерации i_0, чтобы в распределениях правильно выделились малые вероятности, и делаются не на каждой итерации, чтобы модель успевала восстановить адекватность. В экспериментах разреживания включались на итерациях с номерами $i = i_0 + k\delta,\ k = 1, 2, \ldots$, где i_0 и δ — параметры стратегии разреживания.

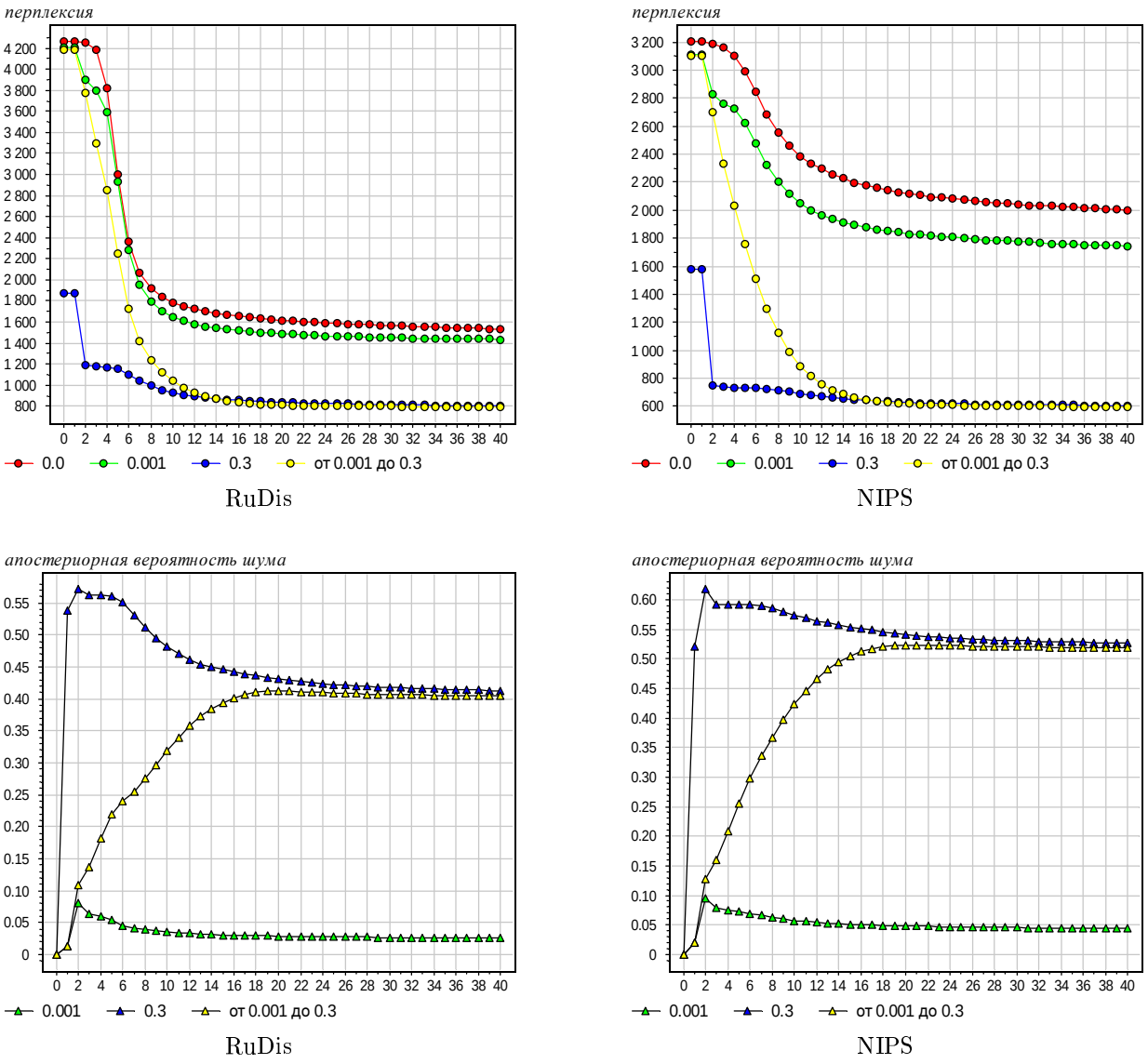


Рис. 5: Зависимость перплексии и апостериорной вероятности шума от числа итераций при $\gamma = 0.0,\ 0.001,\ 0.3$ и при постепенном увеличении γ от 0.001 до 0.3 на первых 20 итерациях. Остальные параметры: $\alpha_t = 0.5,\ \beta_w = 0.01,\ \varepsilon = 0.01,\ |T| = 100$.

Разреживание может приводить к обнулению распределения $p(t \mid d, w)$, тогда термин w интерпретируется как нетематический для документа d. Поэтому разреживание применяется совместно с робастной моделью (15), либо с упрощенной робастной моделью (19).

Результаты экспериментов приведены на рис. 6–8.

При совмещении упрощенной робастной модели, стохастического EM-алгоритма и разреживания достигается наименьшая перплексия и одновременно наибольшая разреженность матрицы Φ — до 99,4% для RuDis и 99,6% для NIPS (см. рис. 6).

В робастных алгоритмах с шумом и фоном разреживание почти не влияет на перплексию и позволяет достигать сопоставимой разреженности (см. рис. 7).

Под «агрессивным» разреживанием понимается уменьшение δ до 1 или уменьшение i_0 до 1 или применение сложной стратегии, когда доля обнуляемых значений не уменьшается с итерациями. При агрессивном разреживании или при использовании стохастического EM-алгоритма возможно разреживание распределений φ_t до 99%. При числе тем $T = 100$ это означает, что каждый термин в среднем относится только к одной теме.

При недостаточном априорном уровне шума $\gamma = 0{,}01$ агрессивное разреживание может приводить к расходимости EM-алгоритма (рис. 8). Тонкие кривые без точек, проходящие чуть ниже кривых контрольной перплексии, соответствуют перплексии на обучающей выборке. Они показывают, что расходимость возникает синхронно на контроле и обучении, причем на обучении расходимость даже более заметна и может быть легко обнаружена во время итераций EM-алгоритма.

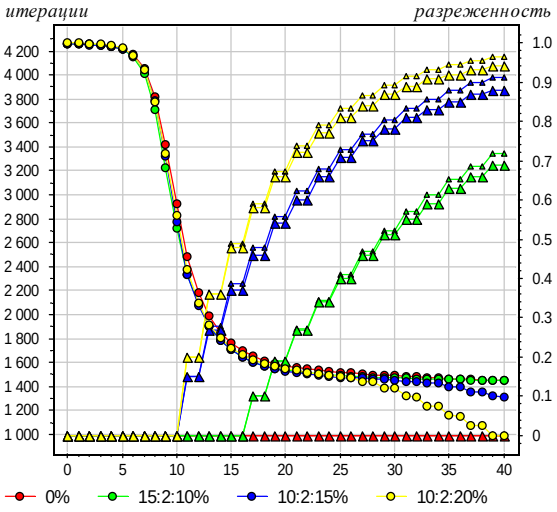
В экспериментах с упрощенной робастной моделью расходимость не наблюдалась.

О дилемме «сглаживание–разреживание»

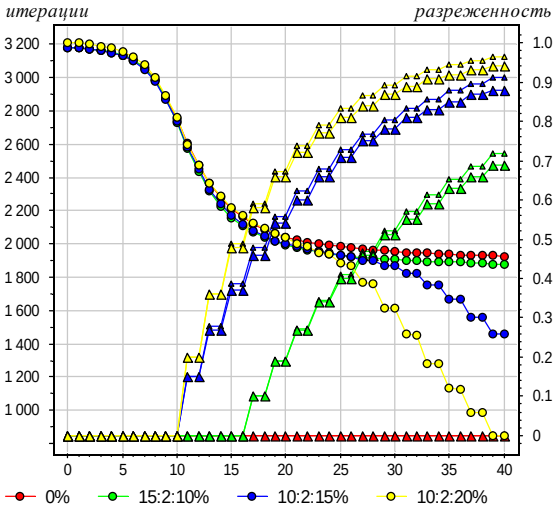
Среди рассмотренных эвристик только сглаживание и разреживание являются взаимно исключающими. В задачах машинного обучения им соответствуют два альтернативных подхода к понижению размерности — отбор признаков и регуляризация. Разреживание является частным случаем отбора признаков, который основан на предположении, что не все признаки несут полезную информацию. Сглаживание является следствием регуляризации, которая основана на предположении, что все признаки полезны и отбрасывать их нельзя, но необходимо ограничить их степени свободы.

В современных исследованиях по тематическому моделированию преобладают методы регуляризации. В данной работе мы показываем, что альтернативный подход к понижению размерности заслуживает не меньшего внимания. При этом разреживание особенно эффективно в сочетании с робастными тематическими моделями.

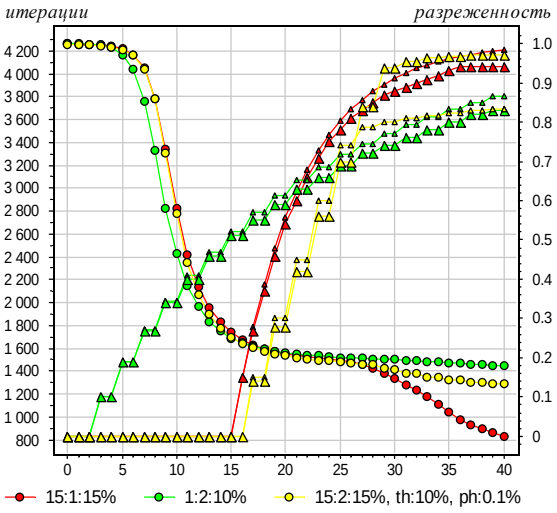
Согласно экспериментам, проведенным в [5], LDA обеспечивает существенно меньшие значения контрольной перплексии, чем PLSA. По аналогии с задачами классификации и регрессии отсюда был сделан стандартный вывод, что модель PLSA имеет слишком много параметров θ_{td}, φ_{wt}, и при отсутствии ограничений на них возникает переобучение. Байесовская регуляризация должна сокращать эффективную размерность и уменьшать переобучение. Однако более тщательное сравнение PLSA и LDA показывает, что регуляризация Дирихле в тематических моделях играет совсем другую роль.



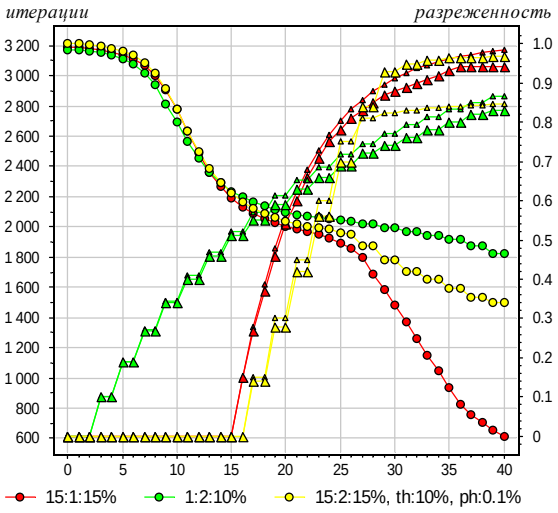
RuDis, разреживание через 2 итерации



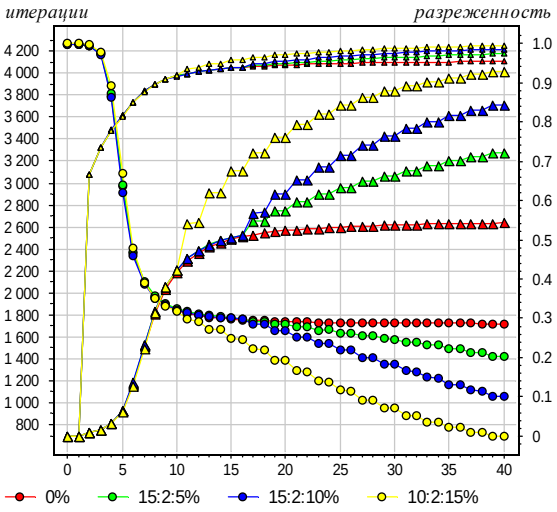
NIPS, разреживание через 2 итерации



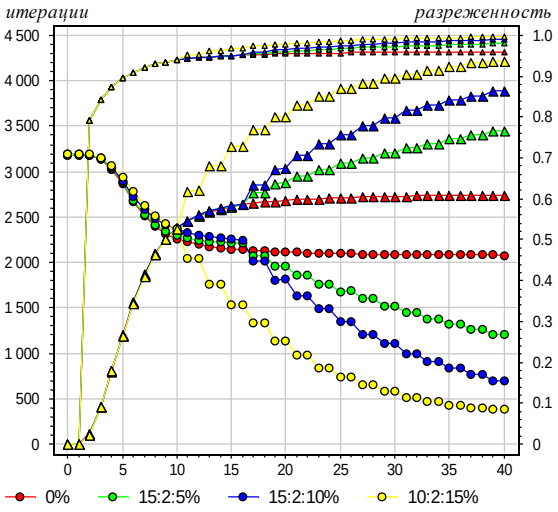
RuDis, агрессивное разреживание



NIPS, агрессивное разреживание

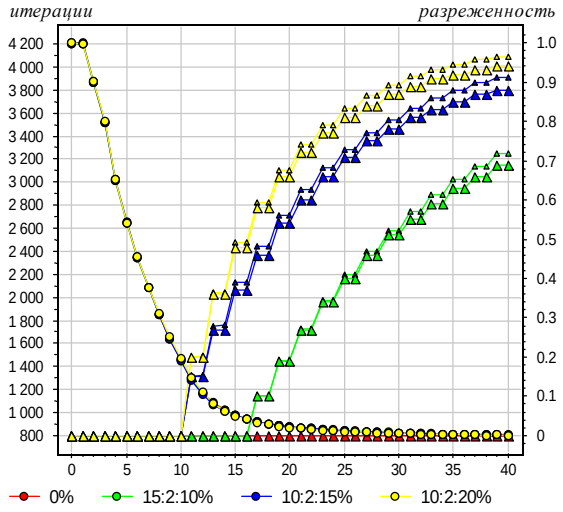


RuDis, SEM, через 2 итерации

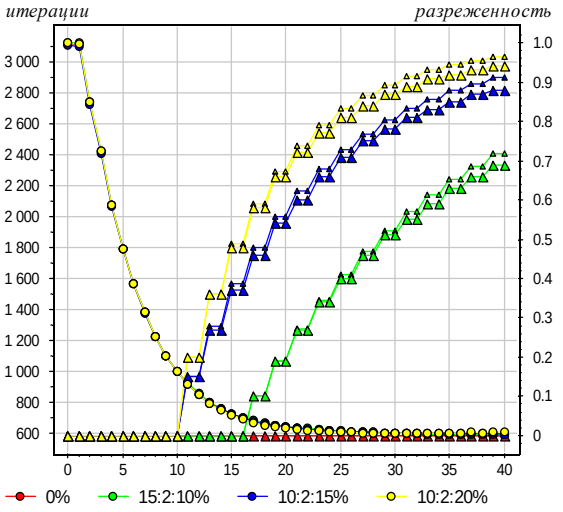


NIPS, SEM, через 2 итерации

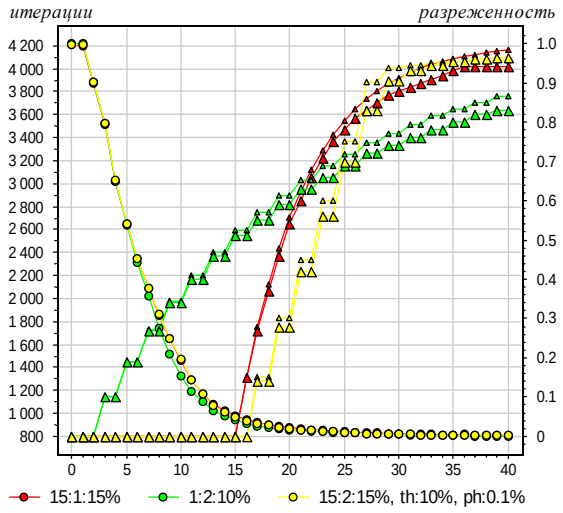
Рис. 6: Зависимость перплексии (○) и разреженности матриц Φ (▵) и Θ (△) от числа итераций для рационального и стохастического EM-алгоритма при различных параметрах разреживания, обозначаемых $i_0{:}\delta{:}r$, $\mathsf{th}{:}R_\theta$, $\mathsf{ph}{:}R_\varphi$. Число тем $|T| = 100$



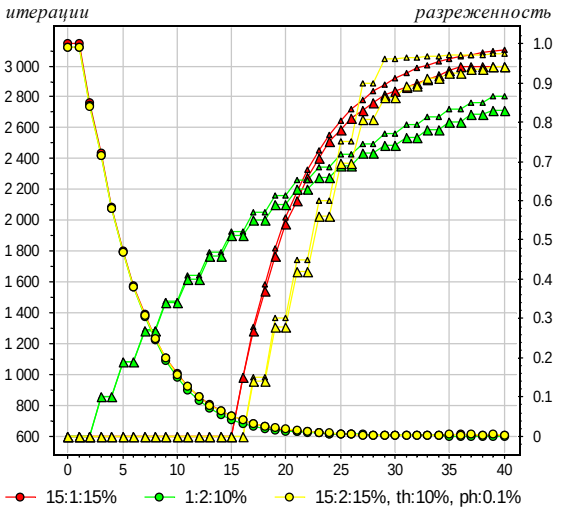
RuDis, разреживание через 2 итерации



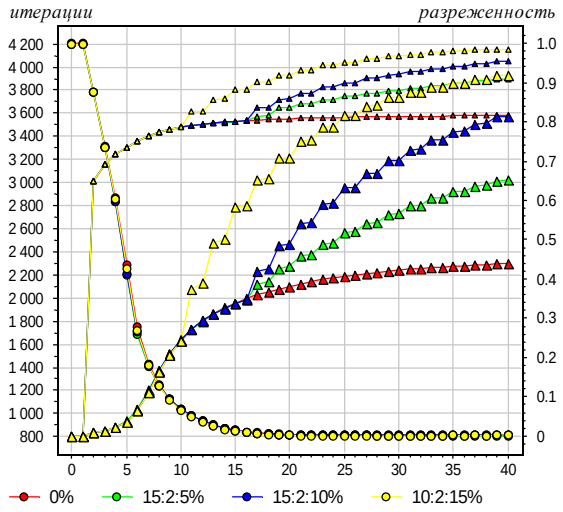
NIPS, разреживание через 2 итерации



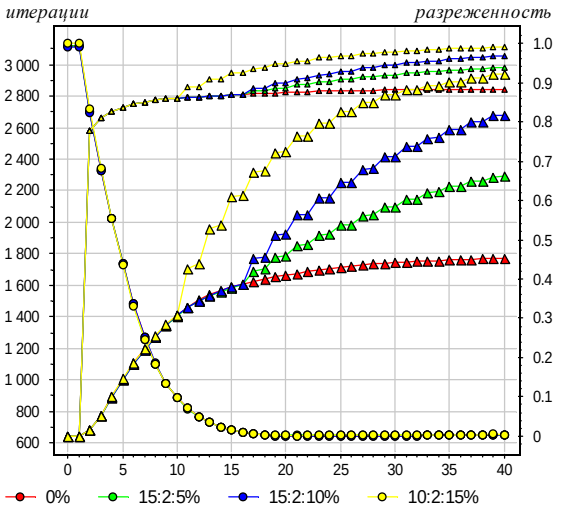
RuDis, агрессивное разреживание



NIPS, агрессивное разреживание

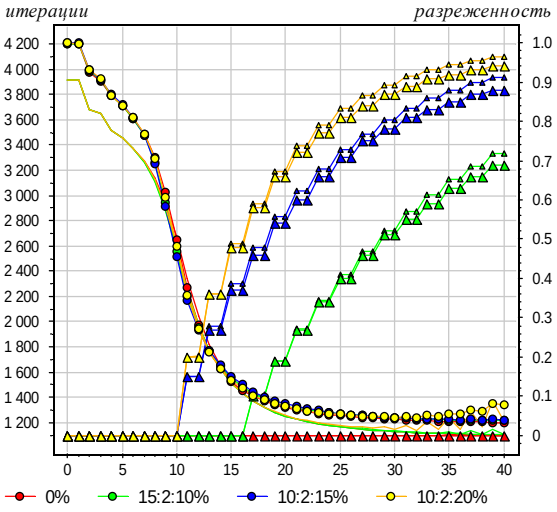


RuDis, SEM, через 2 итерации

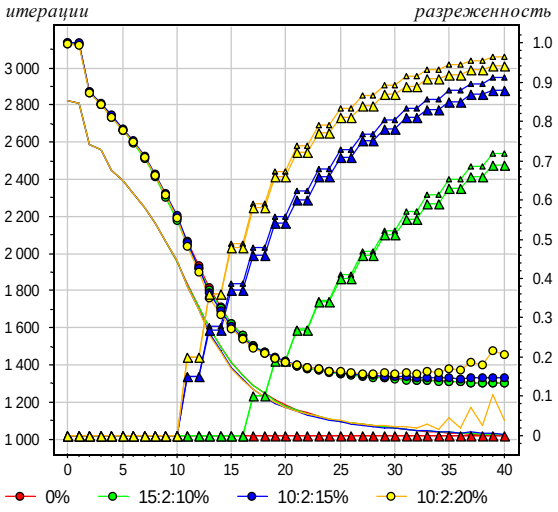


NIPS, SEM, через 2 итерации

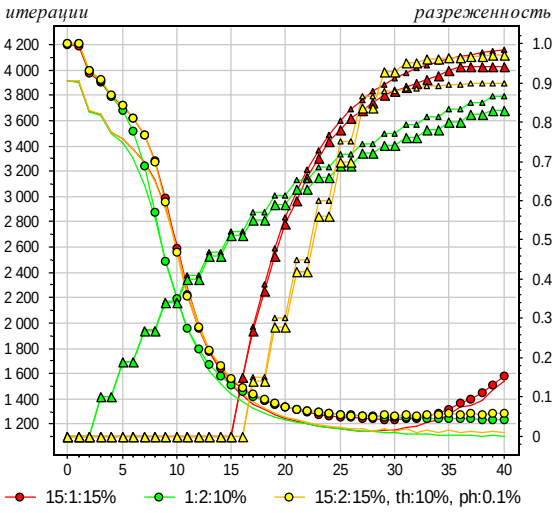
Рис. 7: Зависимость перплексии (○) и разреженности матриц Φ (△) и Θ (△) от числа итераций для рационального и стохастического робастного EM-алгоритма с параметрами робастности $\gamma = 0.3$, $\varepsilon = 0.01$ и параметрами разреживания $i_0{:}\delta{:}r$, $\mathsf{th}{:}R_\theta$, $\mathsf{ph}{:}R_\varphi$. Число тем $|T| = 100$



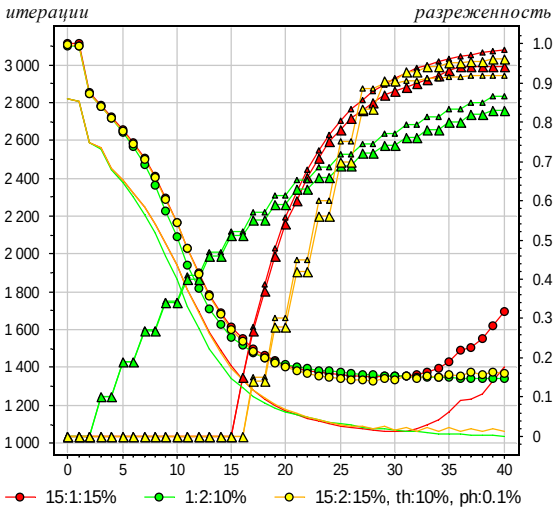
RuDis, разреживание через 2 итерации



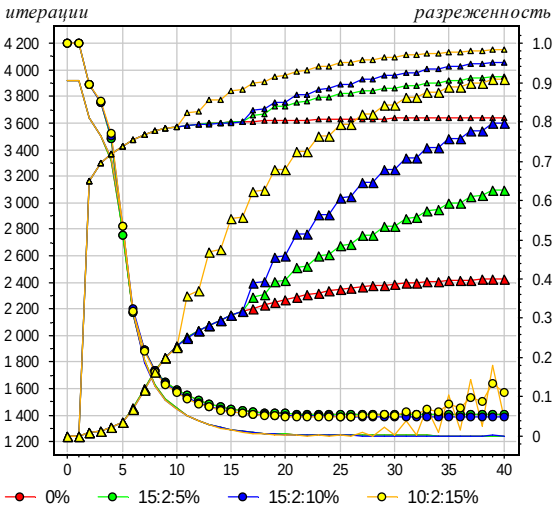
NIPS, разреживание через 2 итерации



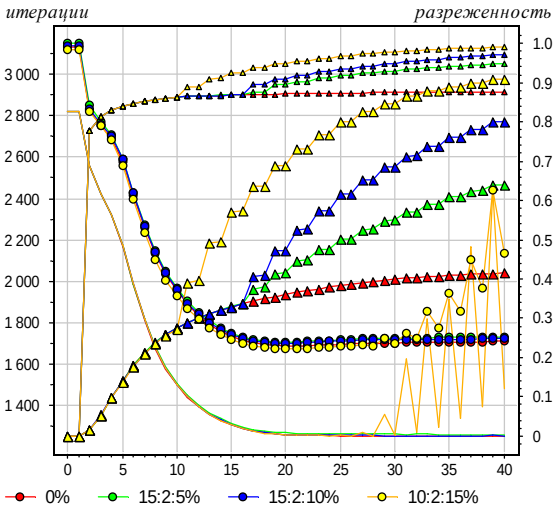
RuDis, агрессивное разреживание



NIPS, агрессивное разреживание



RuDis, SEM, через 2 итерации



NIPS, SEM, через 2 итерации

Рис. 8: Зависимость перплексии (○) и разреженности матриц Φ (△) и Θ (△) от числа итераций для рационального и стохастического робастного EM-алгоритма при малой априорной вероятности шума $\gamma = 0.01$, $\varepsilon = 0.01$, с параметрами разреживания $i_0{:}\delta{:}r$, $\mathsf{th}{:}R_\theta$, $\mathsf{ph}{:}R_\varphi$. Число тем $|T| = 100$

Регуляризация Дирихле приводит к сглаживанию частотных оценок условных вероятностей (13)–(14), что является единственным принципиальным отличием LDA от PLSA. В экспериментах оптимальные значения гиперпараметров α_t и β_w оказываются достаточно близкими к нулю [23]. Для большинства тем в документах $\alpha_t \ll n_{td}$; для большинства терминов в темах $\beta_w \ll n_{wt}$. Оценки параметров φ_{wt} и θ_{td} в PLSA и LDA заметно отличаются только для тем, очень редких в документе, и терминов, очень редких в теме. Они не несут статистически значимой информации о тематике. Их следовало бы проигнорировать, как шум, но вместо этого LDA, наоборот, повышает оценку их вероятности.

Утверждение о том, что LDA сокращает эффективную размерность пространства параметров [5], звучит неубедительно. PLSA и LDA оценивают параметры φ_{wt} и θ_{td} по одним и тем же формулам (13)–(14). Более того, в LDA вводятся дополнительные гиперпараметры α_t, β_w, которые также приходится оценивать [23].

Утверждение о том, что LDA гораздо меньше переобучается [5], не выдерживает аккуратной перепроверки. Качество тематических моделей принято сравнивать по контрольной перплексии, которая может резко повышаться при появлении в контрольных документах редких терминов, для которых модель предсказывает вероятность $p(w \mid d)$, близкую к нулю. В PLSA эта вероятность может оказаться равной нулю, тогда перплексия формально будет равна $+\infty$. Это выглядит как переобучение, однако по сути им не является, так как небольшую долю редких терминов вполне допустимо интерпретировать как нетематический шум. В LDA вероятности редких терминов не стремятся к нулю благодаря сглаженным оценкам φ_{wt} и θ_{td}. Модель LDA более толерантна к нетематическим терминам, но она не выделяет их в явном виде, как это делает робастная модель или SWB.

Если из контрольных документов убрать новые термины, то контрольные перплексии PLSA и LDA практически совпадают [1]. Недавние исследования [16, 26, 15], также подтверждают, что для больших коллекций нет существенных различий в качестве моделей PLSA и LDA. Реальные коллекции настолько велики, что переобучение не является проблемой для обеих моделей. Значимые отличия контрольной перплексии PLSA и LDA в ранних экспериментах [5], могут быть объяснены тем, что для них использовались существенно различные реализации алгоритмов обучения. В наших экспериментах использовался один и тот же алгоритм обучения для моделей PLSA и LDA, отличавшийся только сглаженными оценками в LDA. Сравнивать порождающие модели при существенно различных методах их обучения, вообще говоря, некорректно.

Таким образом, роль априорных распределений Дирихле оказывается весьма скромной — это не сокращение размерности и не уменьшение переобучения, а всего лишь более толерантное оценивание редких терминов, незначимых для выявления тематики. В то же время, регуляризация порождает свои проблемы. Она противоречит гипотезе разреженности и вводит гиперпараметры α_t и β_w, которые приходится подбирать. При появлении документов с новыми терминами w не ясно, как инициализировать β_w. Сглаженные оценки являются смещенными, в отличие от оценок максимума правдоподобия.

Заключение

Описан широкий класс методов тематического моделирования на базе обобщенного EM-алгоритма и эвристик сглаживания, сэмплирования, частого обновления параметров, робастности и разреживания, которые могут сочетаться в различных комбинациях.

В экспериментах на двух текстовых коллекциях получены следующие выводы.

1. Робастные алгоритмы с разреживанием являются лучшими по критерию контрольной перплексии и не требует введения априорных распределений Дирихле. Эвристика сглаживания для них оказывается избыточной.

2. Контрольная перплексия LDA лучше, чем у PLSA не потому, что PLSA переобучается, а потому, что LDA завышает оценки вероятности редких слов. При корректном сравнении на больших коллекциях перплексии PLSA и LDA практически не различаются.

3. Принудительное разреживание в робастных моделях PLSA позволяет обнулять до 99% параметров без ухудшения контрольной перплексии.

4. Упрощенная робастная модель с разреживанием, в отличие от модели SWB, четко выделяет в документах нетематические термины, не требует хранения параметров π_{dw}, не требует задания параметров γ и ε, и почти не увеличивает объем вычислений.

5. Наряду с сэмплированием Гиббса возможны и другие стратегии разреживания распределений $p(t \mid d, w)$, в частности, сэмплирование небольшого фиксированного числа s тем и постепенное разреживание путем обнуления небольшой доли наименьших вероятностей.

6. На достаточно больших коллекциях (10^6 терминов и более) обучающая и контрольная перплексия ведут себя практически одинаково и приводят к одинаковым качественным выводам. Таким образом, нет необходимости вычислять контрольную перплексию.

Работа выполнена при поддержке Российского фонда фундаментальных исследований (проект № 11-07-00480) и программы ОМН РАН «Алгебраические и комбинаторные методы математической кибернетики и информационные системы нового поколения».

Авторы выражают глубокую признательность рецензентам за ценные замечания, способствовавшие улучшению изложения.

Литература

[1] К. В. Воронцов and А. А. Потапенко. Регуляризация, робастность и разреженность вероятностных тематических моделей. *Компьютерные исследования и моделирование*, 4(4):693–706, 2012.

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

[3] С. В. Царьков. Автоматическое выделение ключевых фраз для построения словаря терминов в тематических моделях коллекций текстовых документов. *Естественные и технические науки*, 62(6):456–464, 2012.

[4] A. Asuncion, M. Welling, P. Smyth, and Y. W. Teh. On smoothing and inference for topic models. In *Proceedings of the International Conference on Uncertainty in Artificial Intelligence*, 2009.

[5] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[6] Gilles Celeux, Didier Chauveau, and Jean Diebolt. On stochastic versions of the EM algorithm. Technical Report RR-2514, INRIA, 1995.

[7] C. Chemudugunta, P. Smyth, and M. Steyvers. *Modeling general and specific aspects of documents with a probabilistic topic model*, volume 19, pages 241–248. MIT Press, 2007.

[8] Ali Daud, Juanzi Li, Lizhu Zhou, and Faqir Muhammad. Knowledge discovery through directed probabilistic topic models: a survey. *Frontiers of Computer Science in China*, 4(2):280–301, 2010.

[9] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. of the Royal Statistical Society, Series B*, (34):1–38, 1977.

[10] Jacob Eisenstein, Amr Ahmed, and Eric P. Xing. Sparse additive generative models of text. In *ICML'11*, pages 1041–1048, 2011.

[11] Mark Girolami and Ata Kabán. On an equivalence between PLSI and LDA. In *SIGIR'03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 433–434, 2003.

[12] Thomas Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57, New York, NY, USA, 1999. ACM.

[13] Thomas Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42(1-2):177–196, 2001.

[14] Martin O. Larsson and Johan Ugander. A concave regularization technique for sparse mixture models. In J. Shawe-Taylor, R.S. Zemel, P. Bartlett, F.C.N. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 1890–1898, 2011.

[15] Yue Lu, Qiaozhu Mei, and ChengXiang Zhai. Investigating task performance of probabilistic topic models: an empirical study of PLSA and LDA. *Information Retrieval*, 14(2):178–203, 2011.

[16] Tomonari Masada, Senya Kiyasu, and Sueharu Miyahara. Comparing LDA with pLSI as a dimensionality reduction method in document clustering. In *Proceedings of the 3rd International Conference on Large-scale knowledge resources: construction and application*, LKR'08, pages 13–26. Springer-Verlag, 2008.

[17] Radford M. Neal and Geoffrey E. Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. In Michael I. Jordan, editor, *Learning in graphical models*, pages 355–368. MIT Press, Cambridge, MA, USA, 1999.

[18] Pavel Pecina and Pavel Schlesinger. Combining association measures for collocation extraction. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 651–658. Association for Computational Linguistics, 2006.

[19] A. A. Potapenko and K. V. Vorontsov. Robust PLSA performs better than LDA. In *35th European Conference on Information Retrieval, ECIR-2013, Moscow, Russia, 24-27 March 2013*, pages 784–787. Lecture Notes in Computer Science (LNCS), Springer Verlag-Germany, 2013.

[20] Mark Steyvers and Tom Griffiths. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(Suppl. 1):5228–5235, 2004.

[21] Yee Whye Teh, David Newman, and Max Welling. A collapsed variational bayesian inference algorithm for latent dirichlet allocation. In *NIPS*, pages 1353–1360, 2006.

[22] Hanna Wallach. *Structured Topic Models for Language*. PhD thesis, Newnham College, University of Cambridge, 2008.

[23] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981, 2009.

[24] Chong Wang and David M. Blei. Decoupling sparsity and smoothness in the discrete hierarchical dirichlet process. In *NIPS*, pages 1982–1989. Curran Associates, Inc., 2009.

[25] Yi Wang. Distributed Gibbs sampling of latent dirichlet allocation: The gritty details, 2008.

[26] Yonghui Wu, Yuxin Ding, Xiaolong Wang, and Jun Xu. A comparative study of topic models for topic clustering of chinese web news. In *Computer Science and Information Technology (ICCSIT), 2010 3rd IEEE International Conference on*, volume 5, pages 236–240, july 2010.

[27] Ziqi Zhang, José Iria, Christopher Brewster, and Fabio Ciravegna. A comparative evaluation of term recognition algorithms. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC08)*, 2008.