Московский государственный университет имени М. В. Ломоносова

Факультет Вычислительной Математики и Кибернетики

Кафедра Математических Методов Прогнозирования

ДИПЛОМНАЯ РАБОТА

Лингвистическая регуляризация вероятностных тематических моделей

Выполнил:

студент 5 курса 517 группы

Потапенко Анна Александровна

Научный руководитель:

д.ф-м.н. *Воронцов Константин Вячеславович*

Москва, 2014

Содержание

Аннотация

Вероятностное тематическое моделирование – это современный инструмент статистического анализа текстов, предназначенный для выявления тематики коллекций документов. Стандартные методы находят множество решений, хорошо описывающих коллекцию, однако, не решают задачу выбора наиболее интерпретируемого из них, при котором темы модели напрямую соответствуют темам в человеческом восприятии. Возможный путь состоит в учете требований лингвистического характера, вытекающих из особенностей представления информации в форме текста на естественном языке. В частности, существенная доля слов текста не несет тематическую нагрузку, а служит для связи слов в предложениях, и поэтому должна описываться механизмами, отличными от тематических. При этом темы коллекции должны содержать лишь небольшую долю слов словаря и максимально различаться между собой. Формализация таких требований с помощью аппарата аддитивной регуляризации тематических моделей позволяет построить гибкую модель, которая по результатам экспериментов на реальных данных улучшает интерпретируемость и разреженность тем при незначительном ухудшении перплексии. Интерпретируемость оценивается большим числом показателей, в том числе предлагаются новые меры, показывающие, насколько хорошо выделяется ядро – множество слов, отличающих данную тему от остальных.

1 Введение

Вероятностное тематическое моделирование – статистический аппарат анализа текстов, активно развивающийся на протяжении последних 10 лет. Под темой понимается дискретное распределение на множестве слов, которое показывает, как часто различные слова используются для её описания. Тематическая модель коллекции текстовых документов выявляет представленные в ней темы и описывает каждый документ дискретным распределением на множестве этих тем.

Стандартные тематические модели *вероятностного латентного семантического анализа* (Probabilistic Latent Semantic Analysis, PLSA, [21]) и *латентного размещения Дирихле* (Latent Dirichlet Allocation, LDA, [12]) описывают эмпирические вероятности появления слов в документах модельными распределениями, исходя из принципа максимизации правдоподобия. При этом работают исключительно статистические механизмы, и никак не учитывается лингвистическая природа моделируемых данных.

Это приводит к ряду проблем, в частности, к низкой интерпретируемости получаемых моделей. Предполагается, что человек (эксперт), увидев список наиболее частотных слов и документов темы, должен понять, о чём эта тема, и дать ей адекватное название. Только тогда темы будут полезны для приложений – информационного поиска, категоризации, аннотирования документов. На практике построенные модели часто не отвечают этому требованию. Отдельные темы могут оказаться непонятными, содержать слишком много слов, содержать общеупотребительные слова, казаться смесью нескольких слабо связанных тем, быть слишком похожими на другие темы.

Проанализируем некоторые требования лингвистического характера, нацеленные на повышение адекватности тематической модели.

1. Существенная доля слов не несет тематической нагрузки, а служит лишь для связи слов в предложении. Это могут быть как слова общей лексики, частотные по всей коллекции, так и, наоборот, редкие слова, случайно встретившиеся в данном контексте. Внесение таких слов в основные темы модели зашумляет и снижает их интерпретируемость.

2. Каждая тема характеризуется небольшой долей слов словаря, и каждый документ относится к небольшому числу тем. Таким образом, искомые дискретные распределения слов в темах и тем в документах по своей природе сильно разрежены, т.е. содержат большое число нулей.

3. Темы существенно различны между собой. Каждая тема характеризуется терминологией той предметной области, которой она соответствует. Слова, характерные для некоторой темы, не могут одновременно быть сильно вероятными в большом числе других тем.

4. Тематика меняется плавно вдоль текста. Предложение обычно посвящено одной или двум темам. От предложения к предложению тема меняется редко, от абзаца к абзацу чаще, от раздела к разделу еще чаще.

5. Слова одной темы, часто встречающиеся совместно, могут являются единым термином предметной области, состоящим из нескольких отдельных слов. Целесообразно выделение и внесение таких составных терминов (n-грамм) в словарь модели.

Это лишь начало обширного списка предположений, учет которых позволил бы приблизить тематическую модель к специфике текстов естественного языка и повысить ее качество. Существуют подходы к построению разреженных [35, 23, 48], n-граммных моделей [39, 19, 42]. Однако все они существенно модифицируют модель и алгоритм ее обучения, что затрудняет одновременный учет многих требований.

Целью данной работы является построение гибкой тематической модели, учитывающей набор требований лингвистического характера и улучшающей интерпретируемость тем.

Для построения модели применяется аппарат аддитивной регуляризации [9, 8]. Дополнительные требования формализуются в виде набора регуляризаторов, которые оптимизируются одновременно с правдоподобием модели. Задача тематического моделирования является некорректно поставленной и имеет бесконечно много решений. Известные алгоритмы, такие, как PLSA или LDA, выдают любое из этих решений. Аддитивная регуляризация позволяет направить процесс в сторону максимизации дополнительных критериев и сделать выбор решения более обоснованным.

Многокритериальный подход используется не только на этапе построения, но и на этапе оценивания качества тематической модели. Точность модели измеряется стандартной величиной – *контрольной перплексией*. Контролируются меры разреженности и интерпретируемости модели. Согласно [30, 29] с экспертными оценками интерпретируемости хорошо коррелирует *когерентность*, оценивающая совместную встречаемость наиболее вероятных слов темы во всей коллекции. Помимо когерентности в данной работе используется ряд новых показателей интерпретируемости. Вводится понятие ядра темы – множества характерных терминов, которые с большой вероятностью употребляются в данной теме и практически не встречаются в других темах. Предполагается, что интерпретируемость темы тем лучше, чем больше суммарная вероятность терминов ядра (*чистота темы*) и чем больше вероятность встретить термины ядра именно в данной теме (*контрастность темы*).

Работа организована следующим образом. В разделе 2 приводится краткий обзор стандартных тематических моделей, а также подходов, направленных на учет различных особенностей текстов естественного языка. В разделе 3 вводится и исследуется в экспериментах семейство робастных моделей, использующих дополнительные механизмы для описания нетематических слов коллекции. В разделе 4 предлагаются простые и вычислительно эффективные стратегии разреживания тематических моделей, обеспечивающие до 96% нулей в искомых распределениях без потери точности. В разделе 5 идеи робастности и разреженности моделей обобщаются и дополняются требованиями различности и неизбыточности тем. Предлагается набор регуляризаторов, совместное использование которых повышает интерпретируемость модели.

Эксперименты на коллекции англоязычных статей научной конференции NIPS показывают, что с помощью подходящей комбинации регуляризаторов возможно построить сильно разреженную модель с лучшими показателями интерпретируемости, без значимого ухудшения перплексии модели. Увеличение различности предметных тем приводит к тому, что они очищаются от нетематических слов, частотных по всей коллекции. Выделяемые ядра тем в большинстве случаев удаётся интерпретировать как устоявшуюся терминологию, употребляемую для описания отдельных задач, подходов, методов, отвечающих тематике конференции NIPS.

2 Обзор тематических моделей

2.1 Классические модели PLSA и LDA

Терминология и базовые предположения. Пусть D — множество (коллекция) текстовых документов, W — множество (словарь) всех употребляемых в них терминов. Терминами могут быть как отдельные слова, так и ключевые фразы. 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.

Предполагается, что существует конечное множество тем T, и каждое употребление термина w в каждом документе d связано с некоторой темой $t \in T$, которая не известна. 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 множестве $W \times D \times T$. Термины w и документы d являются наблюдаемыми переменными, тема $t \in T$ является *латентной* (скрытой) переменной.

Предположение о независимости, называемое также гипотезой «мешка слов», означает, что тематику документа можно узнать даже после произвольной перестановки терминов, хотя для человека такой текст теряет смысл. Таким образом, порядок слов не учитывается, и документ представляется как подмножество $d \subset W$, в котором каждому элементу $w \in d$ поставлено в соответствие число n_{dw} вхождений термина w в документ d.

Гипотезой *условной независимости* называется предположение, что появление слов по теме t не зависит от документа: $p(w \mid t) = p(w \mid d, t)$. Согласно формуле полной вероятности и гипотезе условной независимости тематическая модель коллекции представляется в виде:

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

Вероятностная модель (1) описывает порождение коллекции D по известным $p(t \mid d)$ и $p(w \mid t)$. Построение тематической модели — это обратная задача: по известной коллекции D требуется восстановить породившие её дискретные распределения $p(t \mid d)$ и $p(w \mid t)$.

Обычно число тем $|T|$ много меньше $|D|$ и $|W|$, и задачу построения тематической модели можно трактовать как поиск приближённого представления заданной

матрицы частот

$$F = (f_{wd})_{W\times D}, \quad f_{wd} = p(w\,|\,d) = n_{dw}/n_d, \quad f_d = (f_{wd})_{w\in W},$$

в виде произведения $F \approx \Phi\Theta$ двух неизвестных матриц меньшего размера — *матрицы терминов тем* Φ и *матрицы тем документов* Θ:

$$\begin{aligned} \Phi &= (\varphi_{wt})_{W\times T}, \quad & \varphi_{wt} &= p(w\,|\,t), \quad & \varphi_t &= (\varphi_{wt})_{w\in W}; \\ \Theta &= (\theta_{td})_{T\times D}, \quad & \theta_{td} &= p(t\,|\,d), \quad & \theta_d &= (\theta_{td})_{t\in T}. \end{aligned}$$

Матрицы F, Φ, Θ являются *стохастическими*, то есть имеют неотрицательные нормированные столбцы f_d, φ_t, θ_d, представляющие дискретные распределения.

Вероятностный латентный семантический анализ (Probabilistic Latent Semantic Analysis, PLSA) был предложен Томасом Хофманном в [21].

Для построения модели (1) максимизируется логарифм правдоподобия при ограничениях нормировки и неотрицательности:

$$L(\Phi,\Theta) = \ln \prod_{d\in D}\prod_{w\in d} p(w\,|\,d)^{n_{dw}} = \sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{t\in T} \varphi_{wt}\theta_{td} \;\to\; \max_{\Phi,\Theta}; \tag{2}$$

$$\sum_{w\in W} \varphi_{wt} = 1, \quad \varphi_{wt} \geq 0; \qquad \sum_{t\in T} \theta_{td} = 1, \quad \theta_{td} \geq 0. \tag{3}$$

Для решения оптимизационной задачи 2, 3 применяется *EM-алгоритм* – итерационный процесс, в котором каждая итерация состоит из двух шагов — E (expectation) и M (maximization) [15].

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

$$H_{dwt} = p(t\,|\,d,w) = \frac{p(w\,|\,t)p(t\,|\,d)}{p(w\,|\,d)} = \frac{\varphi_{wt}\theta_{td}}{\sum\limits_{s\in T}\varphi_{ws}\theta_{sd}}. \tag{4}$$

На M-шаге, наоборот, при фиксированных вероятностях тем H_{dwt} вычисляются оценки максимального правдоподобия для параметров φ_{wt}, θ_{td}:

$$\varphi_{wt} = \frac{n_{wt}}{n_t}, \qquad n_{wt} = \sum_{d\in D} n_{dw}H_{dwt}, \qquad n_t = \sum_{w\in W} n_{wt}; \tag{5}$$

$$\theta_{td} = \frac{n_{dt}}{n_d}, \qquad n_{dt} = \sum_{w\in d} n_{dw}H_{dwt}, \qquad n_d = \sum_{t\in T} n_{dt}. \tag{6}$$

Начальные приближения φ_t и θ_d в простейшем случае можно задавать нормированными случайными векторами из равномерного распределения.

Алгоритм 2.1 EM-алгоритм для тематической модели PLSA.

Вход: коллекция документов D, число тем $|T|$, начальные приближения Θ и 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 $d \in D,\ w \in d$
    $Z := \sum_{t \in T} \varphi_{wt}\theta_{td}$;
    для всех $t \in T$ таких, что $\varphi_{wt}\theta_{td} > 0$
      увеличить n_{wt}, n_{dt}, n_t на $\delta = n_{dw}\varphi_{wt}\theta_{td}/Z$;
  $\varphi_{wt} := n_{wt}/n_t$ 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 стабилизируются.

Латентное размещение Дирихле. Латентное размещение Дирихле (latent Dirichlet allocation, LDA) предложено Дэвидом Блеем в [12] и является на сегодняшний день доминирующим подходом в вероятностном тематическом моделировании. Модель LDA также основана на разложении (1), однако в целях уменьшения переобучения дополнительно предполагает, 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:

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

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

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

Два наиболее часто используемых метода обучения данной модели основаны на вариационном выводе [36] и на сэмплировании Гиббса [33, 43]. Согласно второму подходу, производится сэмплирование тем для каждой пары $(d,\ w)$ из распределения H_{dwt} и подсчет величин n_{wt}, n_{td}, n_t. Формулы пересчета θ_{td} и φ_{wt} представляют собой сглаженные аналоги формул (5), (6), использовавшихся при обучении PLSA-модели:

$$\varphi_{wt} = \frac{n_{wt} + \beta_w}{n_t + \beta_0}, \quad \beta_0 = \sum_{w \in W} \beta_w; \qquad\qquad \theta_{td} = \frac{n_{dt} + \alpha_t}{n_d + \alpha_0}, \quad \alpha_0 = \sum_{t \in T} \alpha_t. \tag{7}$$

Алгоритм 2.2 LDA-GS: сэмплирование Гиббса для тематической модели LDA.

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

Выход: распределения Θ и Φ;

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

2.2 Разреженные тематические модели

Будем предполагать, что каждый документ относится лишь к небольшому числу тем (если же это энциклопедия, то её лучше разбить на отдельные статьи). Аналогично, 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 $p(t \mid d)$, $p(w \mid t)$, $p(t \mid d, w)$ должны содержать значительную долю нулевых вероятностей. Использование разреженных распределений в модели дает вычислительные преимущества как по памяти, так и по времени работы алгоритмов и является необходимым условием работы с большими объемам текстовых данных.

Модель Fully Sparsed Topic Model [35] строится в тех же предположениях, что и модель PLSA, однако обучающий EM-алгоритм организован иначе.

На *E-шаге* осуществляется оптимизация Θ при фиксированных Φ с помощью алгоритма Франка-Вульфа. Пусть даны профили тем $\varphi_1, \ldots \varphi_{|T|}$ и некоторый документ d. Необходимо найти вектор его латентного представления $\theta_d = (\theta_{1d}, \ldots \theta_{|T|d})$.

По принципу максимума правдоподобия это такой вектор θ_d, что:

$$\ln P(d) = \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \theta_{td} \varphi_{wt} \to \max .$$

Обозначим $x_{dw} = \sum_{t \in T} \theta_{td} \varphi_{wt}$, $x_d = (x_{d1}, \dots x_{d|W|})$. Тогда вектор $x_d = \sum_{t \in T} \theta_{td} \varphi_t$ – это выпуклая линейная комбинация профилей тем $\varphi_1, \dots \varphi_{|T|}$. Таким образом, можно перейти от задачи поиска оптимального вектора θ_d к поиску оптимального вектора x_d на симплексе тем:

$$x_d^* = \arg\max_{x \in \Delta} \sum_{w \in d} n_{dw} \ln x_w,$$

где $\Delta = \text{conv}(\varphi_1, \dots \varphi_{|T|})$.

Это задача вогнутой максимизации на симплексе. Алгоритм Франка-Вульфа ищет разреженное приближение решения такой задачи. Применяя его, получаем алгоритм 2.3. Он обладает следующими свойствами:

1. Алгоритм сходится к оптимальному решению за линейное время.

2. Найденное решение разреженное: после l итераций вектор θ_d имеет не более $l + 1$ ненулевой компоненты.

3. Можно задавать желаемое соотношение между разреженностью решения, близостью решения к оптимальному и временем работы алгоритма: чем больше итераций совершаем, тем больше время работы, тем ближе решение к оптимальному, и тем менее оно разреженное.

На *M-шаге* по коллекции документов и найденным профилям документов θ_{td} оцениваются профили тем φ_{wt}. Задача максимизации правдоподобия приводит к аналитическому решению:

$$\varphi_{wt} \propto \sum_{d \in D} n_{dw} \theta_{td}.$$

Таким образом, матрица тем Φ – это произведение двух разреженных матриц: матрицы встречаемости слов в документах и матрицы латентных представлений документов Θ.

Алгоритм 2.3 E-шаг: алгоритма Франка-Вульфа

Вход: профили тем $\varphi_1, \dots \varphi_{|T|}$, документ d;

Выход: вектор θ_d^*, для которого $\sum\limits_{t\in T} \theta_{td}^* \varphi_t = x_d^*$ максимизирует $f(x_d) = \sum\limits_{w\in d} n_{dw} \ln x_{dw}$;

1: Выбираем вершину φ_r симплекса $\Delta = \text{conv}(\varphi_1, \dots \varphi_{|T|})$ с наибольшим $f(\varphi_r)$:
$x_d^0 := \varphi_r, \quad \theta_{rd}^0 := 1, \quad \theta_{kd}^0 := 0, \ \forall k \neq r;$

2: **для** $l = 1, \dots, \infty$

3: $\quad i' := \arg\max\limits_i \varphi_i^T \nabla f(x_d^l);$

4: $\quad \alpha' = \arg\max\limits_{\alpha\in[0,1]} f(\alpha\, \varphi_{i'} + (1-\alpha)\, x_d^l);$

5: $\quad x_d^{l+1} := \alpha'\, \varphi_{i'} + (1-\alpha')\, x_d^l;$

6: $\quad \theta_d^{l+1} := (1-\alpha')\theta_d^l; \quad \theta_{i'd}^{l+1} := \theta_{i'd}^{l+1} + \alpha';$

Регуляризация задачи приближенного матричного разложения. Будем интерпретировать задачу построения тематической модели как задачу приближенного матричного разложения $F \approx \Phi\Theta$. Начнем с более простых подходов, которые не учитывают условия нормировки на матрицы F, Φ, Θ.

В [14, 45] рассматривается задача латентного семантического анализа в опмтимизационной формулировке:

$$\|F - \Phi\Theta\|^2 \to \min_{\Phi,\Theta}, \quad s.t.\ \Phi^T\Phi = I, \tag{8}$$

где $\|\cdot\|$ – евклидова норма.

В [14] для получения разреженной матрицы Φ используется l_1-регуляризация:

$$\|F - \Phi\Theta\|^2 + \lambda\|\Phi\|_1 \to \min_{\Phi,\Theta}, \quad s.t.\ \Phi^T\Phi = I, \tag{9}$$

где $\|\cdot\|_1$ – $l1$-норма. Параметр λ контролирует степень разреженности матрицы Φ.

Матрицы Φ и Θ могут быть найдены итерационным алгоритмом. При фиксированной Θ задача (9) распадается на $|W|$ независимых задач по строкам матриц, каждая из которых решается покоординатным спуском. При фиксированной Φ решение ищется с помощью SVD-разложения. В задачу (9) может быть дополнительно добавлено требование неотрицательности элементов матрицы Φ, схема решения остается прежней.

В [45] вводится модель регуляризованного латентного семантического анализа (Regularized Latent Semantic Indexing, RLSI), согласно которой предлагается использовать l_2-регуляризатор для матрицы Φ и l_1-регуляризатор для матрицы Θ.

Недостатком этих методов является использование евклидовой нормы, т.к. принцип максимума правдоподобия соответствует минимизации дивергенции Кульбака-Лейблера. В [48] для нахождения латентого представления больших коллекций данных предлагается так называемое разреженное тематическое кодирование (Sparse Topical Coding, STC). Согласно данному подходу матрица Φ сохраняется и имеет привычный вероятностный смысл. Для каждого слова в документе вводится *код слова* – ненормированный вектор $s_{dw} \in \mathbb{R}^{|T|}$, определяющий вес каждой темы. Для каждого документа вводится *код документа* – ненормированный вектор $\theta_d \in \mathbb{R}^{|T|}$, который может быть найден некоторой агрегацией кодов входящих в документ слов. Задача ставится как минимизация ненормированной дивергенции Кульбака-Леблера между наблюдаемыми счетчиками n_{dw} вхождений слов в документы и их представлением в виде $s_{dw}^T \varphi_t$. Для достижения разреженных кодов документов θ_d в оптимизируемый функционал добавляется l_1-нормы векторов θ_d.

Основная трудность в учете условия нормировки на матрицы F, Φ и Θ при поиске разреженного разложения $F \approx \Phi\Theta$ состоит в том, что использование l_1-регуляризации оказывается бесполезным: все матрицы имеют фиксированную l_1-норму. Тем не менее, в [23] предлагается способ обойти эту проблему. Рассмотрим задачу минимизации расстояния Кульбака-Леблера между матрицами F и $\Phi\Theta$:

$$\sum_{i=1}^{|W|}\sum_{j=1}^{|D|} F_{ij} \log \frac{F_{ij}}{(\Phi\Theta)_{ij}} + (\Phi\Theta)_{ij} \to \min_{\Phi,\Theta}, \quad s.t. \Phi, \Theta \geq 0 \tag{10}$$

Можно показать, что задача (10) для нормированной матрицы F эквивалентна задаче максимизации правдоподобия, решаемой при выводе PLSA модели.

Для получения разреженной матрицы Θ снова добавим регуляризатор, однако вместо l_1-нормы будем использовать сумму логарифмов: $\sum_{ij} \log(\Theta_{ij} + \varepsilon)$, где ε – параметр. Решение такой регуляризованной задачи оптимизации эквивалентно MAP оценкам при выборе в качестве априорного распределения на профили документов θ_d так называемого псевдо-Дирихле распределения. Его плотность задается соотношением:

$$p(x_1, \dots x_N) = C(\alpha, \varepsilon) \prod_{i=1}^{N} (x_i + \varepsilon_i)^{\alpha_i - 1}, \quad x \in \Delta^{N-1}, \tag{11}$$

где $\alpha \in \mathbb{R}^N, \varepsilon \in \mathbb{R}_+^N$ – параметры распределения, $C(\alpha, \varepsilon)$ – нормировочная константа.

Это аналог распределения Дирихле. Отличие состоит в том, что оно имеет ограниченную плотность, в то время как распределение Дирихле не ограничено в случае $\alpha < 1$. Именно эта область значений соответствует эффекту разреживания, и следовательно, представляет особый интерес. Неограниченность функции плотности вызывает трудности при выводе оценок. Введение распределения (11) решает эти проблемы.

Алгоритм обучения модели представляет собой EM-алгоритм для PLSA с заменой формулы (6) обновления θ_{td} на формулу:

$$\theta_{td} = \frac{n_{dt}}{n_d + (1-\alpha)\left(\frac{1}{\varepsilon+\theta_{td}} - \sum_{t' \in T} \frac{\theta_{t'd}}{\varepsilon+\theta_{t'd}}\right)} \tag{12}$$

с последующей нормировкой.

Данный подход может быть естественным образом обобщен и на матрицу Ф.

Построение разреженных профилей тем. Другой подход к достижению разреженности заключается в использовании более сложной порождающей модели, которая бы лучше учитывала особенности естественного языка. Так, в [44] предполагается, что каждая тема описывает не все слова словаря, а лишь некоторое их подмножество. Вводятся дополнительные бинарные переменные b_{wt} из распределения Бернулли, которые для каждой пары слово-тема определяют, относится ли данное слово к данной теме. Каждая тема t описывается распределением Дирихле на подмножестве слов, заданном переменными b_{wt}. Таким образом, профили тем содержат сглаженные оценки, но при этом разрежены. Сглаженность и разреженность регулируется независимо параметрами распределения Дирихле и распределения Бернулли. Обучение модели проводится с помощью схемы Гиббса с дополнительными шагами для оценки параметров b_{wt}. Недостатком данной модели является большое число дополнительных скрытых переменных, которые усложняют обучение.

2.3 Выделение нетематических слов

В текстах коллекции лишь небольшая доля слов являются терминами и характеризуют тематику документа. Остальные слова служат для связи слов в предложениях, являются частотными словами общей лексики или, наоборот, употребляются

крайне редко. Темы модели необходимо освободить от таких слов, отфильтровав их и описав вероятности их появления другими механизмами, отличными от тематических.

Введение дополнительной фоновой компоненты. В [16] вводится некоторое фоновое вероятностное распределение на всем словаре, а распределение каждой темы задается как отклонение от него:

$$p(w \mid t) \propto \exp(\eta_t + m),$$

где $m \in \mathbb{R}^{|W|}$ определяет фоновое распределение, а $\eta_t \in \mathbb{R}^{|W|}$ характеризует конкретную тему. При таком подходе вместо «переменных-переключателей» используется простое сложение вероятностных распределений в логарифмической шкале. При этом η_t является разреженным вектором и содержит ненулевые значения только для слов, отличающих данную тему от остальных. В фоновом распределении, наоборот, большую вероятность получают общеупотребительные слова, которые встречаются в документах любой тематики.

В некоторых моделях [13] вводится не только фоновое распределение, общее для всей коллекции, но и дополнительные распределения, свои для каждого документа, которые описывают редкие случайные термины или особенности стиля автора.

Учет слововхождений с весами. Очистить тематическую модель от нетематических слов можно не вводя дополнительных компонент в явном виде, а лишь уменьшив вклад таких слов в тематическую модель. В [46] предлагается взвешивание терминов согласно следующим схемам.

В теории информации, если вероятность события a равна $p(a)$, то количество информации – это $-\log_2 p(a)$. Рассматривая термины как события, можем найти количество информации, содержащееся в термине w, или, другими словами, вес термина:

$$m(w) = -\log_2 p(w) = -\log_2 \frac{n_w}{n},$$

где n_w – число вхождений термина w в коллекцию, n – длина коллекции.

Чтобы учесть термины пропорционально их весам предлагается модифицировать формулу сэмплирования Гиббса, используя вместо счетчиков вхождений терминов

их суммарный вес:

$$p(t|d,w) = \frac{m(w)n_{wt} + \beta}{\sum_{w'} m(w')n_{w't} + W\beta} \frac{\sum_{w'} m(w')n_{dw't} + \alpha}{\sum_{w'} m(w')n_{dw'} + T\alpha}$$

Тогда каждое слово w документа d, отнесенное к теме t, дает вклад $m(w)$, а не 1.

В рассмотренной схеме вес зависит только от самого термина. В более сложной схеме Pointwise Mutual Information вес термина может быть разным в разных документах:

$$m(w,d) = \log_2 \frac{p(w|d)}{p(w)}$$

Данная модель использовалась для задачи поиска по документу его перевода на другом языке и привела к более высокой точности сопоставления, чем стандартная модель LDA. Поиск перевода осуществлялся из условия минимизации дивергенции Йенсена-Шеннона между тематическими профилями документов.

Скрытая марковская модель функциональных слов. В [18] производится отказ от гипотезы «мешка слов», документы рассматриваются как последовательности слов, и моделирование функциональных нетематических слов языка осуществляется с помощью скрытой марковской модели (Hidden Markov Model, HMM).

Предполагается, что каждое слово в тексте имеет синтаксическую роль, описываемую HMM, и небольшое подмножество слов имеет также семантическую роль, описываемую LDA. Строится гибрид: скрытая марковская модель, у которой в одном выделенном классе слова генерируются согласно тематической модели. Обучение проводится на основе полного байесовского вывода и схемы Гиббса. При этом существенно расширяется множество параметров модели, которое включает вероятности перехода между классами скрытой марковской модели, вероятности слов в рамках каждого класса для HMM и стандартные матрицы Φ и Θ для тематической модели, в которых не учитываются слова, ушедшие в нетематические классы.

В результате разделение слов по классам HMM оказывается очень близким с разделением слов по частям речи. В выделенном тематическом классе группируются существительные, при этом сильно частотные существительные могут выделяться в отдельные нетематические классы модели.

2.4 Отказ от гипотезы «мешка слов»

Предположение о том, что порядок слов в документе не важен, сильно нереалистично. Отказ от этого предположения может быть полезен не только для более детального описания функциональных слов текста, но и для построения более точных тематических моделей, различным образом учитывающих грамматическую и семантическую зависимость соседних слов в предложении.

Скрытая марковская модель на темах слововхождений. Модель Hidden Markov Topic Model (HMTM, [20]) формализует гипотезу о том, что тематика меняется плавно, и тема слововхождения часто сохраняется для следующего. Снова строится скрытая марковская модель текста, однако теперь ее классами являются темы. С определенной вероятностью происходит переход из темы в неё же, с оставшейся вероятностью тема меняется, и тогда новая тема генерируется согласно стандартной модели LDA. Модель обучается с помощью EM-алгоритма, включающего в себя модификацию алгоритма «вперед-назад», часто используемого при работе со скрытой марковской моделью.

N-граммные тематические модели позволяют характеризовать тему не только отдельными словами, но и терминами, состоящими из нескольких слов. Биграммная тематическая модель (Bigram Topic Model, BTM) [39] является обобщением модели LDA и работает с биграммами – парами соседних слов. Аналогичная модель может быть построена и на базе PLSA [11]. Для каждой темы вместо одного распределения $p(w|t)$ вводится W распределений $p(w|v,t)$, отражающих вероятность появления слова w для данной темы t при условии того, что предыдущее слово было v. Таким образом, матрица Φ приобретает размерность $W \times W\,T$.

Основной недостаток модели BTM заключается в том, что она строится только на биграммах, при этом никакие слова не рассматриваются как униграммы. В [19] представлена более гибкая модель LDA Collocation Model (LDACOL), решающая эту проблему. В ней вводятся бинарные величины x_{vw}, которые для каждого слова w в документе указывают, составляет ли данное слово с предыдущим словом v биграмму. К параметрам модели добавляется матрица Ψ размерности $W \times 2$, содержащая вероятности продолжить данное слово биграммой, и матрица Σ размерности $W \times W$,

содержащая вероятности закончить биграмму словом w, если она была начата словом v.

Слово в документе порождается следующим образом. Сначала определяется, образуется ли биграмма с предыдущим словом v согласно вероятностям из Ψ. Затем генерируется тема для позиции из профиля документа θ_d Если биграмма образуется, то генерируется слово с вероятностями σ_v для известного предыдущего слова v, если нет – то с вероятностями φ_t для данной темы t.

Недостатком данной модели является то, что никак не определяются тематики биграмм. Темы, по-прежнему характеризуются отдельными словами, а биграммы задаются вероятностями перехода от слова к слову, вне зависимости от темы.

Модель Topical N-gramm Model (TNG) [42] наследует от BTM независимую обработку коллокаций в рамках различных тем и от LDACOL умение генерировать и биграммы, и униграммы. Матрицы Ψ и Σ из предыдущей модели получают дополнительную размерность по темам T. Слово в документе порождается следующим образом. Сначала определяется, образуется ли биграмма с предыдущим словом согласно вероятностям ψ_{vs} для предыдущего слова v и от его темы s. Затем генерируется тема для позиции из профиля документа θ_d. Если биграмма образуется, то генерируется слово с вероятностями σ_{vt} для известного предыдущего слова v и данной темы t, если нет – то с вероятностями φ_t для данной темы t.

Обучение всех трех моделей проводится на основе схемы Гиббса, появляются дополнительные счетчики, соответствующие соседним слововхождениям и их тематикам. Все модели обобщаются на случай n-грамм при $n > 2$.

Предварительное выделение n-грамм. Рассмотренные модели существенно увеличивают число переменных и плохо применяются на практике, особенно при работе с большими объемами данных. В [24] предлагается разделить процесс на два независимых этапа: (1) – выделение биграмм с помощью t-критерия Стьюдента, (2) – построение стандартной тематической модели по мешку отдельных слов и выделенных биграмм. Такой подход не усложняет модель и является более эффективным с вычислительной точки зрения.

Согласно экспериментам на нескольких текстовых коллекциях выделение биграмм позволяет улучшить качество модели сразу по трем существенно различным

показателям: критерию AIC $= -2L(\Phi, \Theta) + |W||T|$, основанному на правдоподобии модели, когерентности [30], оценивающей интерпретируемость тем, и качеству классификации документов на коллекции с известной разметкой. При этом существует оптимальное количество выделенных биграмм, отклонение от которого приводит к ухудшению результатов.

3 Робастные тематические модели

3.1 Робастная модель с шумом и фоном

Робастная тематическая модель формализует предположение, что лишь некоторые слова в текстах относятся к каким-либо темам. Она представляет собой вероятностную смесь трёх компонент — тематической, шумовой и фоновой [1, 2]:

$$p(w \mid d) = \frac{Z_{dw} + \gamma\pi_{dw} + \varepsilon\pi_w}{1 + \gamma + \varepsilon}; \qquad Z_{dw} = \sum_{t \in T} \varphi_{wt}\theta_{td}. \tag{13}$$

Шумовая компонента $\pi_{dw} \equiv p_{\text{ш}}(w \mid d)$ — 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 $\varphi_{wt} = p(w \mid t)$, 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 $\frac{1}{1+\gamma+\varepsilon}$, $\frac{\gamma}{1+\gamma+\varepsilon}$, $\frac{\varepsilon}{1+\gamma+\varepsilon}$ соответственно.

Похожая модель SWB (special words with background) на основе LDA и сэмплирования Гиббса предлагалась в [13].

Задача максимизации правдоподобия (2) для модели (13) решена в [1]. По аналогии со стандартным EM-алгоритмом, на E-шаге для каждой пары (d, w) вычисляются

по формуле Байеса условные вероятности тем $H_{dwt} = p(t \mid d, w)$,

$$H_{dwt} = \frac{\varphi_{wt}\theta_{td}}{Z_{dw} + \gamma\pi_{dw} + \varepsilon\pi_w}, \quad t \in T, \tag{14}$$

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

$$H_{dw} = \frac{\gamma\pi_{dw}}{Z_{dw} + \gamma\pi_{dw} + \varepsilon\pi_w}; \qquad H'_{dw} = \frac{\varepsilon\pi_w}{Z_{dw} + \gamma\pi_{dw} + \varepsilon\pi_w}. \tag{15}$$

На М-шаге переменные θ_{td} и φ_{wt} вычисляются по прежним формулам (5) и (6) с единственным отличием, что теперь H_{dwt} вычисляются по новой формуле (14). Переменные π_{dw} и π_w вычисляются как частотные оценки условных вероятностей шума и фона:

$$\pi_{dw} = \frac{\nu_{dw}}{\nu_d}, \qquad \nu_{dw} = n_{dw}H_{dw}, \qquad \nu_d = \sum_{w \in d} \nu_{dw}, \tag{16}$$

$$\pi_w = \frac{\nu'_w}{\nu'}, \qquad \nu'_w = \sum_{d \in D} n_{dw}H'_{dw}, \qquad \nu' = \sum_{w \in W} \nu'_w, \tag{17}$$

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 (2), 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 $\pi_{dw} > 0$.

3.2 Эксперименты

Коллекции текстов. Вычислительные эксперименты данного раздела проводились на двух коллекциях.

Коллекция RuDis содержит $|D| = 2000$ авторефератов диссертаций на русском языке; суммарная длина составляет $n \approx 8.7 \cdot 10^6$, объём словаря $|W| \approx 3 \cdot 10^4$. Контрольная коллекция D' состоит из 200 авторефератов. Предварительно производилась лемматизация и отбрасывались стоп-слова.

Коллекция NIPS содержит $|D| = 1566$ текстов статей научной конференции Neural Information Processing Systems на английском языке; суммарная длина составляет $n \approx 2.3 \cdot 10^6$, объём словаря $|W| \approx 1.3 \cdot 10^4$. Контрольная коллекция D' состоит из 174 документов. Предварительная обработка текстов включала приведение к нижнему регистру, удаление пунктуации и удаление стоп-слов с помощью библиотеки BOW toolkit [28].

Оцениваение качества. Точность описания тематической моделью $p(w \,|\, d)$ коллекции документов D принято оценивать с помощью *перплексии*:

$$\mathscr{P}(D, p) = \exp\Big(-\frac{1}{n} L(\Phi, \Theta)\Big) = \exp\bigg(-\frac{1}{n} \sum_{d \in D} \sum_{w \in d} n_{dw} \ln p(w \,|\, d)\bigg). \tag{19}$$

Контрольная перплексия $\mathscr{P}(D', p_D)$ вычисляется по контрольной выборке документов D' для модели p_D, построенной по обучающей выборке документов D, не пересекающейся с D'. В наших экспериментах использовалось случайное разбиение коллекции в пропорции $|D| : |D'| = 9 : 1$. Каждый контрольный документ d разбивался случайным образом на две половины: по первой оценивались параметры θ_d, по второй вычислялась перплексия. Параметры φ_t оценивались только по обучающей коллекции. Для робастной модели параметры π_w оценивались по обучающей выборке D, параметры ν_d оценивались по первой каждого документа, параметры π_{dw} оценивались для каждой пары (d, w) согласно (18). Если в контрольных докментах оказывались термины, которых не было в обучающей коллекции D, то они игнорировались.

Обобщенное семейство алгоритмов обучения. EM-алгоритм 2.1 максимизации правдоподобия для модели PLSA и алгоритм 2.2 обучения модели LDA, основанный на сэмплировании Гиббса, выведены в сильно разных предположениях и с использованием разных математических аппаратов, однако имеют много общего на алгоритмическом уровне. Оба алгоритма являются итерационными процессами,

которые многократно просматривают коллекцию документов, пересчитывая вероятности тем для каждого слововхождения и значения искомых распределений Φ и Θ.

При этом ключевые различия заключаются в трех независящих друг от друга опциях: *сэмлпирования* (S), *частого обновления параметров* и *сглаживания* (D). Далее подробно рассматривается влияние каждой из них на обучение модели, и строится семейство моделей с их произвольной комбинацией. Робастная модель рассматривается как еще одна опция (R), которая дополняет обобщенную модель шумовой и фоновой компонентами. В экспериментах исследуется совместимость и взаимодействие всех эвристик.

1. Сэмплирование. На шаге 6 алгоритма 2.1 производится увеличение всех счетчиков пропорционально вероятностям тем для слова w в документе d согласно апостериорному распределению $p(t \mid d, w)$. Будем обозначать такой вариант как P. Вместо этого может производиться сэмплирование s случайных тем t_{dwi}, $i = 1, \ldots, s$ из того же распределения и его замена несмещенной эмпирической оценкой по сгенерированной случайной выборке:

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

Такая модификация называется *стохастическим* EM-алгоритмом, его сходимость обеспечивается несмещенностью оценок φ_{wt}, θ_{td}.

Объём s сэмплируемых выборок является параметром метода. Чем меньше число s, тем более разрежена эмпирическая оценка распределения $p(t \mid d, w)$. В алгоритме LDA-GS число s различно для различных пар (d, w) и равно счетчику n_{dw}. Эксперименты показывают, что достаточно сэмплировать совсем небольшое число тем, около 5 тем обычно достаточно, см. таблицы 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 чуть менее чувствительна к выбору параметра экономного сэмплирования s, см. таблицу 2.

Таблица 1: Стохастический EM-алгоритм для модели LDA. Зависимость перплексии на обучении и контроле от объёма 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 объёма 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

2. Частое обновление параметров. Известно, что в EM-алгоритме нет необходимости слишком точно решать задачу максимизации правдоподобия на каждом M-шаге. Достаточно сместиться в направлении максимума и затем выполнить E-шаг. Модификация EM-алгоритма, состоящая в более частом выполнении E-шага, на-

зывается *обобщённым EM-алгоритмом* (generalized EM-algorithm, GEM). Для него остаются справедливы те же обоснования сходимости, что и для основного варианта EM-алгоритма [15].

В случае PLSA обобщённый EM-алгоритм приводит к более частому обновлению параметров θ_{td} и φ_{wt} по значениям счётчиков n_{wt} и n_{dt}. В Алгоритме 2.1 обновления происходят после каждого прохода коллекции (шаги 7, 8). Возможны варианты обновлений: после каждого документа, после каждого термина (d, w), после заданного числа терминов, после каждого вхождения термина.

Эксперименты показывают, что частота обновления влияет на скорость сходимости и почти не влияет на значение контрольной перплексии в конце итераций, рис. 1.

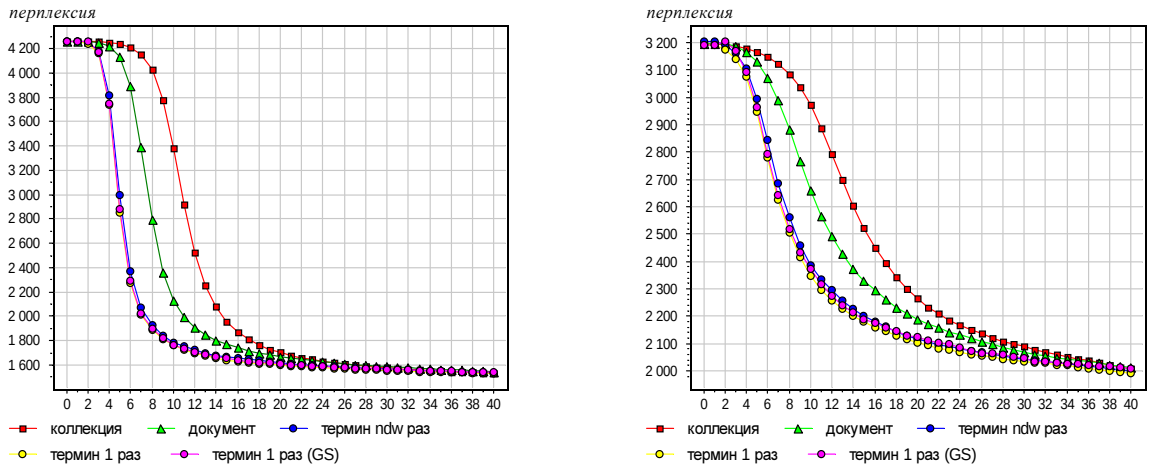
В LDA-GS параметры φ_{wt} и θ_{td} обновляются предельно часто — после обработки каждого вхождения термина w в документ d. Кроме того, перед сэмплированием счётчики уменьшаются на единицу (шаг 5). Тем самым при оценивании распределений не учитывается i-е вхождение термина w в документ d, для которого сэмплируется тема t_{dwi}. Из теории следует, что эта особенность исключительно важна [43]. 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. 1 практически совпадают.

3. Сглаживание. Опция сглаживания соответствует замене формул для вычисления параметров Φ и Θ на шагах 7, 8 алгоритма 2.1 их сглаженными аналогами (7), которые получаются при предположении об априорном распределении Дирихле.

На рис. 2 представлено сравнение восьми алгоритмов, образуемых всеми комбинациями эвристик сглаживания, робастности и сэмплирования, которое позволяет сделать следующие выводы:

1. Сэмплирование работает немного хуже пропорционального учета. Сглаживание улучшает перплексию алгоритмов, использующих сэмплирование.

2. Робастность снижает перплексию эффективнее, чем сглаживание. Для робастных алгоритмов сглаживание не требуется.



RuDis NIPS

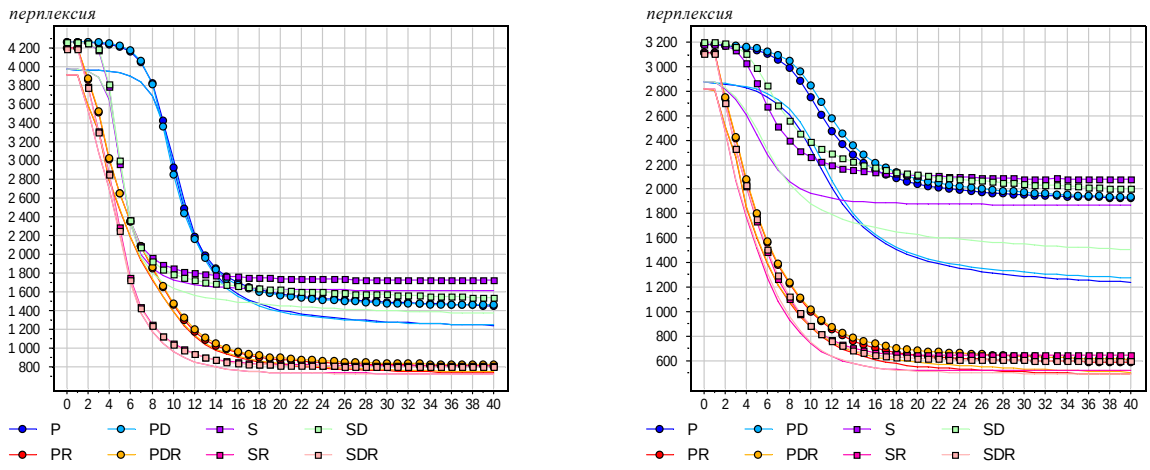
Рис. 1: Зависимость перплексии от числа итераций в стохастическом EM-алгоритме (SEM) при различной частоте обновления параметров φ_{wt}, θ_{td}: после каждого прохода коллекции, после каждого документа, после каждого термина (d, w) по всем n_{dw} его вхождениям, после каждого вхождения термина, GS — с предварительным уменьшением счётчиков как в алгоритме сэмплирования Гиббса. Параметры сглаживания: $\alpha_t = 0.5$, $\beta_w = 0.01$. Число тем $|T| = 100$.

4. Робастность. Из рис. 2 видно, что для обеих задач робастные алгоритмы существенно превосходят неробастные и гораздо меньше переобучаются. Остановимся подробнее на деталях реализации робастности.

Возможны два варианта реализации М-шага — мультипликативный (16), (17) и аддитивный (18). 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 «жёстко» приписывается одна из трёх взаимоисключающих ролей. В экспериментах эти два варианта также не дают значимых различий перплексии.

Зависимость перплексии от параметров γ и ε, как правило, монотонная, причём параметр γ гораздо сильнее влияет на перплексию, чем ε, см. таблицу 3. С ростом γ перплексия уменьшается, так как компонента шума близка к униграммной модели



RuDis NIPS

Рис. 2: 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; $|T| = 100$. 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, $\gamma = 0.3$:

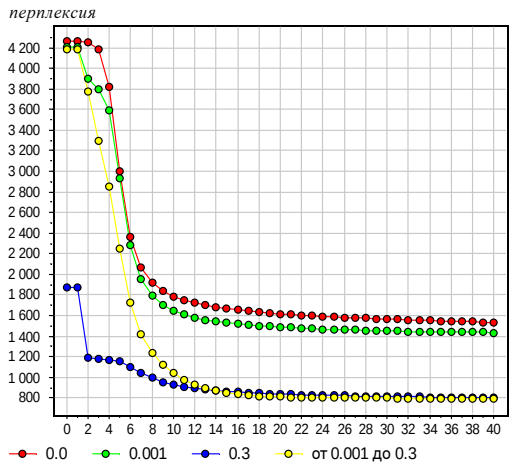
ε	$\mathscr{P}$	$\hat{p}_{\text{ф}}$
0	797	0.000
0.01	794	0.006
0.05	798	0.027
0.1	809	0.049
0.2	823	0.086
0.3	841	0.116
0.5	870	0.165

NIPS, $\varepsilon = 0.01$:

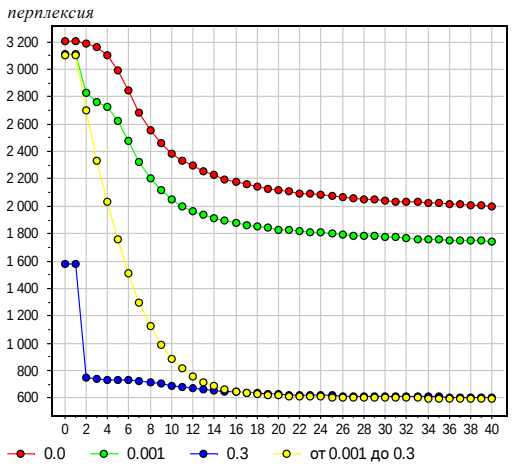
γ	$\mathscr{P}$	$\hat{p}_{\text{ш}}$
0	2001	0.000
0.001	1763	0.044
0.01	1381	0.152
0.05	991	0.296
0.1	818	0.377
0.3	604	0.527
0.5	525	0.598

NIPS, $\gamma = 0.3$:

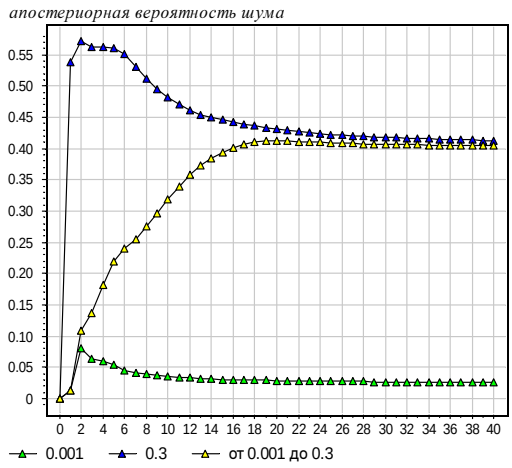
ε	$\mathscr{P}$	$\hat{p}_{\text{ф}}$
0	598	0.000
0.01	596	0.005
0.05	605	0.023
0.1	613	0.043
0.2	630	0.079
0.3	640	0.109
0.5	668	0.157



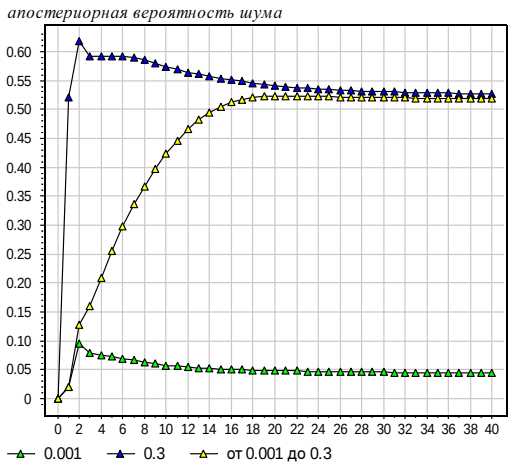
RuDis



NIPS



RuDis



NIPS

Рис. 3: 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: $\alpha_t = 0.5,\ \beta_w = 0.01,\ \varepsilon = 0.01,\ |T| = 100$.

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. 3 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 ещё не успела сойтись.

4 Разреживание тематических моделей

Рассмотренные в обзоре способы построения разреженных тематических моделей приводят к существенному усложнению как порождающей модели, так и алгоритма обучения. Простой и понятной идеей кажется постепенное обнуление наименьших значений вероятностей в ходе итерацией EM-алгоритма и получение таким образом разреженных распределений. Приведем далее развитие и математические обоснование этой идеи и предложим конкретные стратегии разреживания для тематической модели PLSA.

4.1 Подход к разреживанию на основе метода OBD/OBS

Допустим, что EM-алгоритм сошелся в точку локального максимума правдоподобия $L(\Phi, \Theta)$, и первые производные правдоподобия (2) по всем параметрам φ_{wt},

θ_{td} равны нулю. Зададимся вопросом: обнуление каких параметров меньше всего повлияет на значение правдоподобия? Применим ту же технику, которая используется в методе оптимального разреживания многослойных нейронных сетей OBD (Optimal Brain Damage) [25]. Разложив правдоподобие в ряд Тейлора в окрестности точки максимума, получим квадратичную форму по приращениям параметров $\Delta\varphi_{wt}$, $\Delta\theta_{td}$:

$$L(\Phi+\Delta\Phi,\Theta+\Delta\Theta) \approx L(\Phi,\Theta) + \frac{1}{2}\sum_{w,t}\sum_{u,s}\Delta\varphi_{wt}\Delta\varphi_{us}\frac{\partial^2 L(\Phi,\Theta)}{\partial\varphi_{wt}\partial\varphi_{us}} + \frac{1}{2}\sum_{t,d}\sum_{s,g}\Delta\theta_{td}\Delta\theta_{sg}\frac{\partial^2 L(\Phi,\Theta)}{\partial\theta_{td}\partial\theta_{sg}} + \sum_{w,t}\sum_{s,d}\Delta\varphi_{wt}\Delta\theta_{sd}\frac{\partial^2 L(\Phi,\Theta)}{\partial\varphi_{wt}\partial\theta_{sd}}$$

Выпишем производные первого порядка:

$$\frac{\partial L(\Phi,\Theta)}{\partial\varphi_{wt}} = \sum_d n_{dw}\frac{\theta_{td}}{p(w\mid d)}, \quad \frac{\partial L(\Phi,\Theta)}{\partial\theta_{td}} = \sum_w n_{dw}\frac{\varphi_{wt}}{p(w\mid d)}$$

и ненулевые производные второго порядка:

$$\frac{\partial^2 L(\Phi,\Theta)}{\partial\varphi_{wt}\partial\varphi_{ws}} = -\sum_d n_{dw}\frac{\theta_{td}\theta_{sd}}{p^2(w\mid d)},$$

$$\frac{\partial^2 L(\Phi,\Theta)}{\partial\theta_{td}\partial\theta_{sd}} = -\sum_w n_{dw}\frac{\varphi_{wt}\varphi_{ws}}{p^2(w\mid d)},$$

$$\frac{\partial^2 L(\Phi,\Theta)}{\partial\varphi_{wt}\partial\theta_{sd}} = n_{dw}\left(\frac{[t=s]}{p(w\mid d)} - \frac{\varphi_{ws}\theta_{td}}{p^2(w\mid d)}\right)$$

Заметим, что в стандартном методе OBD обычно пренебрегают смешанными частными производными, чтобы упростить вид квадратичной формы. В данном случае увидим, что делать какие-либо приближения нет необходимости.

Обнулить какой-то параметр φ_{wt} означает положить $\varphi_{wt}+\Delta\varphi_{wt}=0$, откуда следует $\Delta\varphi_{wt}=-\varphi_{wt}$. Аналогично, $\Delta\theta_{td}=-\theta_{td}$.

Рассмотрим три случая обнулений:

- Пусть обнуляется один параметр φ_{wt}. Тогда изменение функционала оценивается как:

$$s_{wt} = L(\Phi+\Delta\Phi,\Theta+\Delta\Theta)-L(\Phi,\Theta) = -\frac{1}{2}\varphi_{wt}^2\sum_d n_{dw}\frac{\theta_{td}^2}{p(w\mid d)^2} = -\frac{1}{2}\sum_d n_{dw}H_{dwt}^2$$

- Обнулим теперь параметры φ_{wt} для некоторого фиксированного w. Это эквивалентно выбрасыванию слова из словаря. Тогда получим:

$$s_w = L(\Phi+\Delta\Phi,\Theta+\Delta\Theta)-L(\Phi,\Theta) = -\frac{1}{2}\sum_{d,t,s} n_{dw}\frac{\varphi_{wt}\theta_{td}}{p(w\mid d)}\frac{\varphi_{ws}\theta_{sd}}{p(w\mid d)} = -\frac{1}{2}\sum_d n_d w = -\frac{n_w}{2}$$

- Наконец, рассмотрим обнуление параметров φ_{wt}, θ_{td} для всех слов, тем и документов. Получим выражение:

$$s = L(\Phi + \Delta\Phi, \Theta + \Delta\Theta) - L(\Phi, \Theta) = -\frac{1}{2}\sum_{w,t,s}\varphi_{wt}\varphi_{ws}\sum_{d} n_{dw}\frac{\theta_{td}\theta_{sd}}{p(w \mid d)^2}$$
$$-\frac{1}{2}\sum_{d,t,s}\theta_{td}\theta_{sd}\sum_{w} n_{dw}\frac{\varphi_{wt}\varphi_{ws}}{p(w \mid d)^2} + \sum_{w,t}\sum_{s,d}\varphi_{wt}\theta_{sd}n_{dw}\left(\frac{[t=s]}{p(w \mid d)} - \frac{\varphi_{ws}\theta_{td}}{p^2(w \mid d)}\right) \tag{21}$$

Третья сумма обнуляется, первые две упрощаются к виду:

$$s = L(\Phi + \Delta\Phi, \Theta + \Delta\Theta) - L(\Phi, \Theta) = -\frac{1}{2}\sum_{wt} n_{wt} - \frac{1}{2}\sum_{td} n_{td}$$

Из полученных формул можно сделать вывод о том, что значения счетчиков n_{wt}, n_{td} вносят аддитивные вклады в изменение правдоподобия ΔL. Это обосновывает интуитивно очевидную стратегию разреживания: после каждого прохода коллекции в каждом распределении $\varphi_{wt} = n_{wt}/n_t$ и $\theta_{td} = n_{dt}/n_d$ обнуляются наименьшие значения вероятностей, для которых сумма счётчиков n_{wt} и n_{dt}, соответственно, не превышает заданный порог. Кроме того, логично рассмотреть и другие близкие стратегии: обнуление по пороговому значению вероятности, обнуление определенной доли ненулевых элементов, обнуление определенного числа элементов и различные комбинации этих критериев.

4.2 Упрощенная робастная модель

Обнуление существенной доли элементов в профилях тем φ_{wt} и документов θ_{td} может приводить к обнулениям в распределении $p(w \mid d)$. С одной стороны, ситуация интерпретируется как вполне нормальная: модель может не предсказывать появления слова в документе, например, потому что это слово является своего рода «шумом». Отбрасывание таких слов полезно с вычислительной точки зрения, так как сокращается эффективная длина документа, и проход по нему осуществляется быстрее. Однако последствия таких обнулений оказываются катастрофическими для вероятностной модели: в функционале правдоподобия (2) под логарифмом появляется нуль, перплексия (19) обращается в бесконечность, распределение тем (4) для данного слова не существует.

Это противоречие препятствует достижению сильной доли разреженности в стандартной модели PLSA. Чтобы этого избежать, предлагается скорректировать модель (1) введением дополнительной шумовой компоненты, которая включается только тогда, когда слово в документе не относится моделью ни к одной теме, т.е. величина

$$Z_{dw} = \sum_{t \in T} \varphi_{wt} \theta_{td}$$

обращается в 0. Тогда порождающая модель примет вид:

$$p(w \mid d) = \nu_d \sum_{t \in T} \varphi_{wt} \theta_{td} + \big[Z_{dw} = 0\big] \pi_{dw}, \tag{22}$$

где π_{dw} — новые параметры модели, $\pi_{dw} > 0$ тогда и только тогда, когда $Z_{dw} = 0$; нормировочный множитель ν_d выбирается из условия $\sum\limits_{w \in d} p(w \mid d) = 1$.

Обозначим $\beta_{dw} = \big[Z_{dw} > 0\big]$, $\bar{\beta}_{dw} = \big[Z_{dw} = 0\big]$. Запишем задачу максимизации правдоподобия:

$$L(\Phi, \Theta, \Pi) = \sum_{d \in D, w \in d} n_{dw} \beta_{dw} \log Z_{dw} + \sum_{d \in D} \log \nu_d \sum_{w \in d} n_{dw} \beta_{dw} + \sum_{d \in D, w \in d} n_{dw} \bar{\beta}_{dw} \log \pi_{dw} \to \max_{\Phi, \Theta, \Pi}. \tag{23}$$

Обозначив $n'_{dw} = n_{dw} \beta_{dw}$ получаем задачу на Φ, Θ, абсолютно аналогичную рассматриваемой при выводе PLSA. Формулы также получаются аналогичными, меняется n_{dw} на n'_{dw}:

$$\varphi_{wt} = \frac{\sum\limits_{d \in D} n'_{dw} H_{dwt}}{\sum\limits_{d \in D, w \in w} n'_{dw} H_{dwt}}, \quad \theta_{td} = \frac{\sum\limits_{w \in d} n'_{dw} H_{dwt}}{\sum\limits_{w \in d, t \in T} n'_{dw} H_{dwt}}, \quad H_{dwt} = \frac{\varphi_{wt} \theta_{td}}{\sum\limits_{s \in T} \varphi_{ws} \theta_{sd}}. \tag{24}$$

Таким образом, при вычислении параметров φ_{wt} и θ_{td} все нетематические термины просто игнорируются. Найдем параметры π_{dw} и ν_d. Для этого продифференцируем функционал L по π_{dw} с учетом зависимости

$$\nu_d = 1 - \sum_{w \in d} \big[Z_{dw} = 0\big] \pi_{dw} \tag{25}$$

и приравняем производную нулю:

$$\frac{\partial L}{\partial \pi_{dw}} = -\frac{\sum\limits_{w \in d} n'_{dw}}{\nu_d} \bar{\beta}_{dw} + n_{dw} \bar{\beta}_{dw} \frac{1}{\pi_{dw}} = 0, \quad \text{откуда } \pi_{dw} = \frac{\nu_d n_{dw}}{\sum\limits_{w \in d} n'_{dw}}. \tag{26}$$

Поставляя (26) в зависимость (25), получаем итоговые формулы:

$$\nu_d = \frac{\sum\limits_{w \in d} n'_{dw}}{n_d}, \quad \pi_{dw} = \frac{n_{dw}}{n_d}.$$

Оценка на π_{dw} для всех нетематических терминов совпадает с частотной оценкой условной вероятности $p(w \mid d)$ в униграммной модели. Нормировочный множитель ν_d равен доле тематических терминов в документе. Заметим, что для вычисления матриц Φ и Θ параметры π_{dw} и ν_d вообще не нужны. Они могут понадобиться только для расчёта $p(w \mid d)$ в перплексии модели (22).

Такая коррекция модели представляет собой упрощенный вариант рассмотренной ранее робастной модели. Как и прежде, вводится шумовое распределение, однако доля шумовых слов не выбирается априорно равной некоторому числу (γ), а настраивается автоматически в результате обнуления тематической компоненты. Упрощённая робастная модель не требует хранения дополнительных параметров π_{dw}, число которых сопоставимо с размером коллекции, и не увеличивает время обучения модели.

4.3 Эксперименты

Применим описанный подход построения разреженных тематических моделей на практике. Все модели будем сравнивать по контрольной перплексии и достигаемой разреженности: доле нулевых элементов в матрице Φ и в матрице Θ. Цель – достижение максимальной разреженности при неухудшении перплексии.

Сравниваемые стратегии. *Простая стратегия*: в каждом из распределений φ_t, θ_d обнуляется заданная доля r наименьших *ненулевых* значений. После обнуления производится перенормировка распределений. Поскольку доля берётся от числа ненулевых значений, число обнуляемых значений постепенно сокращается от итерации к итерации. Обнуления прекращаются, когда в распределении остаётся $\lfloor r^{-1} \rfloor$ ненулевых значений. Недостатком этой стратегии является «выровненность» 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 не более заданной доли r наименьших значений, но так, чтобы сумма

обнуляемых вероятностей не превышала заданного порога R_φ для распределений φ_t и заданного порога R_θ для распределений θ_d.

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.

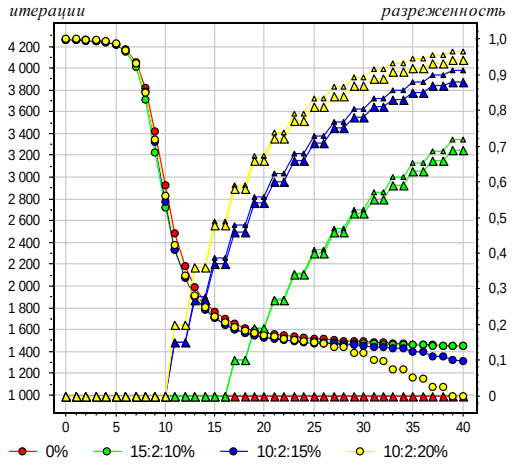
Алгоритмы, подвергаемые разреживанию. Будем применять описанные стратегии постепенного разреживания к различным алгоритмам введенного обобщенного семейства. Для нерегуляризованных алгоритмов с пропорциональным учетом вероятностей тем на E-шаге (P) и с сэмплированием Гиббса (S) будем использовать скорректированную модель (22) и назвать их неробастными несмотря на аналогию используемой корректировки с робастной моделью.

Такое же разреживание профилей проведем для робастных моделей семейства (PR, SR). Согласно формуле (13) при обнулении тематической компоненты Z_{dw} вероятность $p(w \mid d)$ остается положительной за счета шума и фона, поэтому модель не требует дополнительной корректировки.

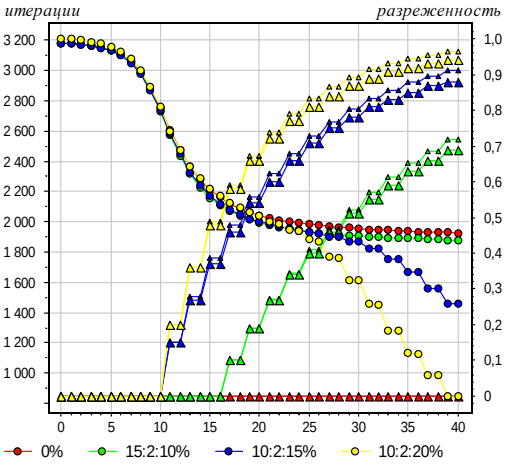
Результаты. Результаты экспериментов приведены на рис. 4, 5, 6. Под «агрессивным» разреживанием понимается либо уменьшение δ до 1, либо уменьшение i_0 до 1, либо применение сложной стратегии, когда число обнуляемых значений не уменьшается с итерациями.

Как в неробастных алгоритмах, так и в робастных удается обнулить более 90% вероятностей в распределениях φ_t, θ_d (рис. 4) без потери качества. В неробастных алгоритмах при использовании разреживания перплексия уменьшается. Это связано с переходом слов из тематической компоненты в шумовую согласно скорректированной модели (22). Робастные алгоритмы имеют более низкую перплексию, при разреживании она не изменяется.

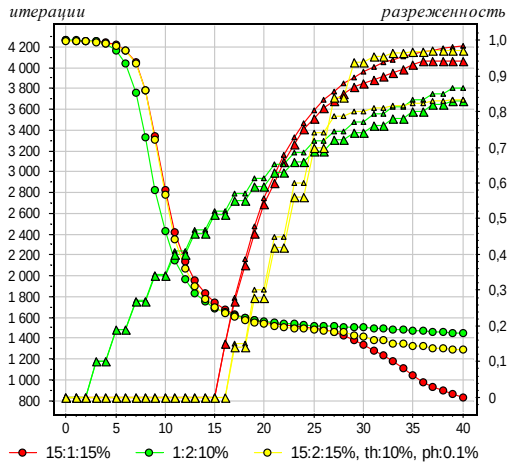
Наиболее сильная (до 96%) одновременная разреженность распределений φ_t, θ_d достигается на обеих коллекциях робастным алгоритмом PR со сложной стратегией разреживания при $i_0 = 15,\ \delta = 2,\ r = 0.15,\ R_\theta = 0.1,\ R_\varphi = 0.001$.



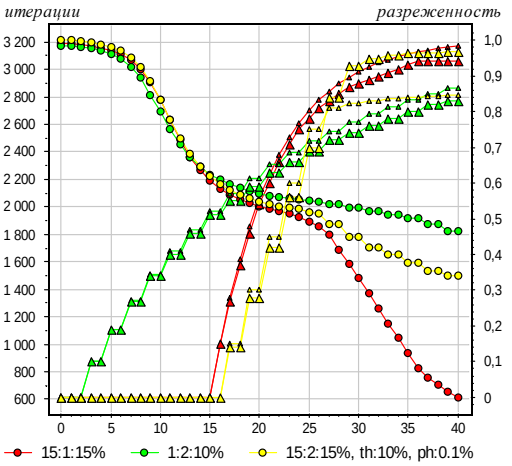
RuDis, P, разреживание через 2 итерации



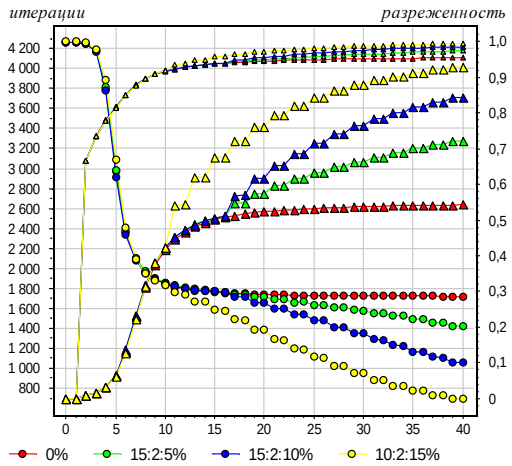
NIPS, P, разреживание через 2 итерации



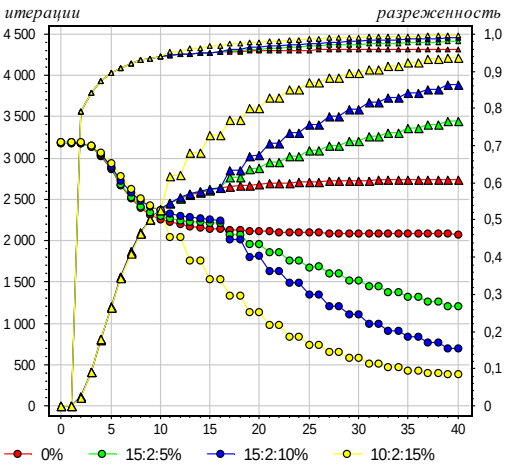
RuDis, P, агрессивное разреживание



NIPS, P, агрессивное разреживание

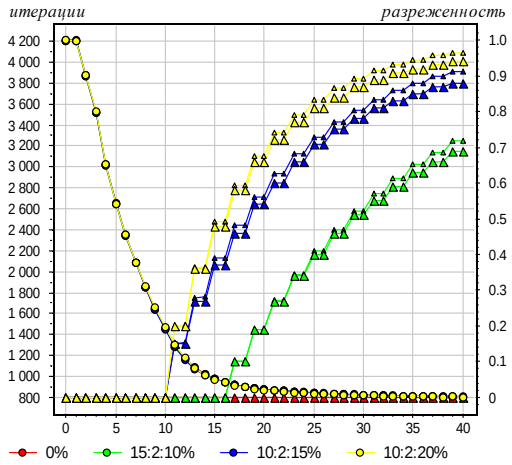


RuDis, S, разреживание через 2 итерации

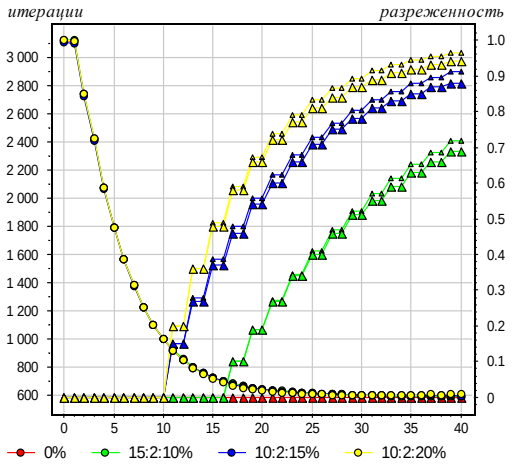


NIPS, S, разреживание через 2 итерации

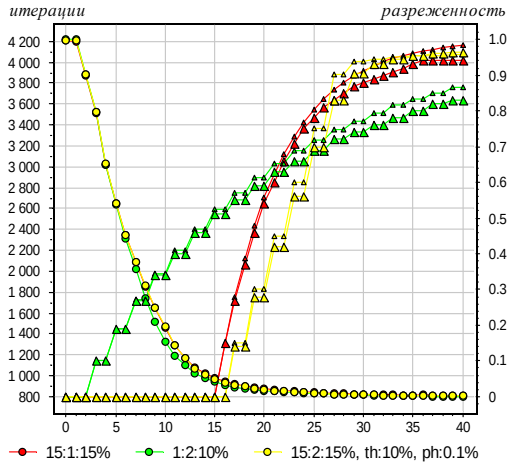
Рис. 4: Зависимость перплексии (○) и разреженности матриц Φ (△) и Θ (△) от числа итераций при разреживании $p(t \mid d)$, $p(w \mid t)$ Параметры разреживания обозначаются $i_0{:}\delta{:}r$, $\mathsf{th}{:}R_\theta$, $\mathsf{ph}{:}R_\varphi$.



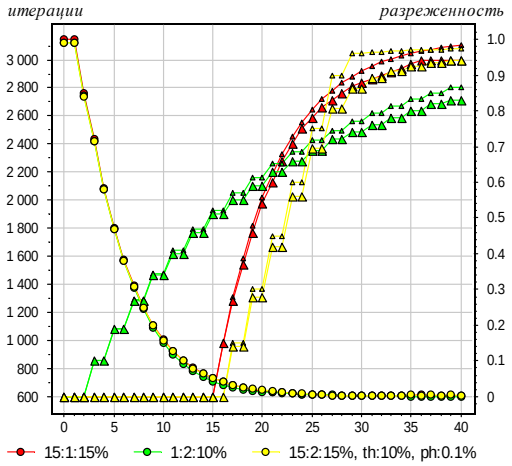
RuDis, PR, разреживание через 2 итерации



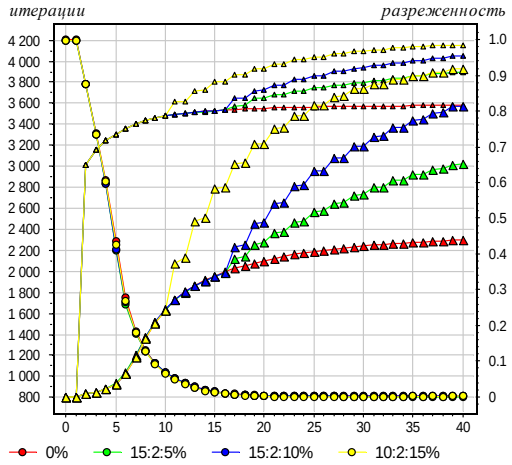
NIPS, PR, разреживание через 2 итерации



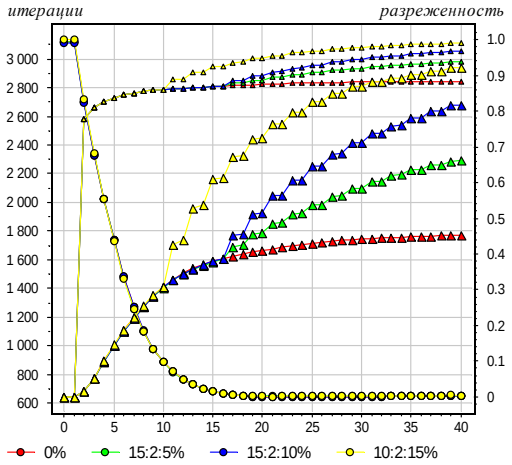
RuDis, PR, агрессивное разреживание



NIPS, PR, агрессивное разреживание

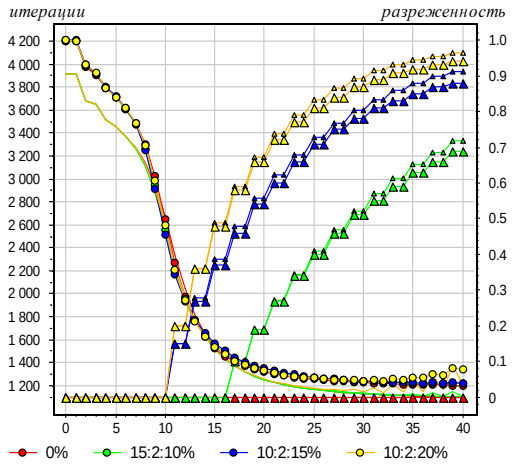


RuDis, SR, разреживание через 2 итерации

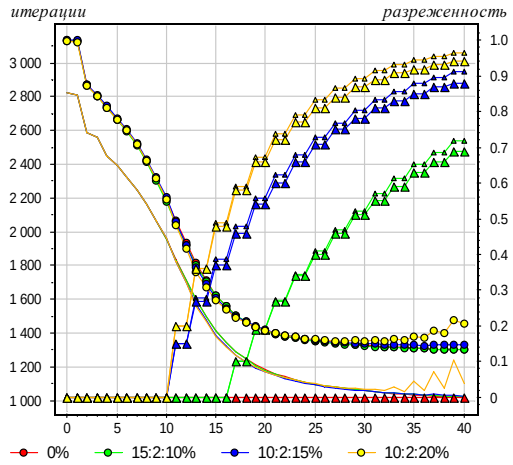


NIPS, SR, разреживание через 2 итерации

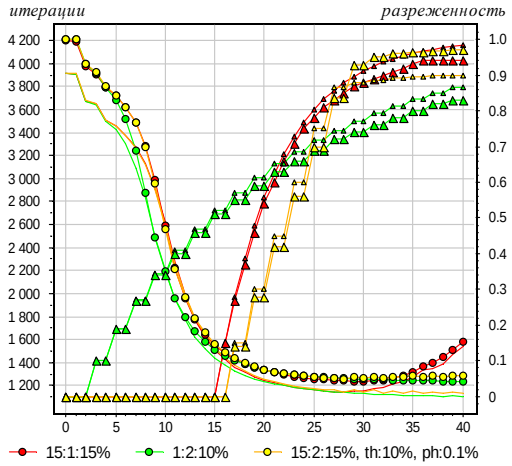
Рис. 5: Зависимость перплексии (○) и разреженности матриц Φ (△) и Θ (△) от числа итераций при разреживании $p(t \mid d)$, $p(w \mid t)$. Параметры разреживания $i_0{:}\delta{:}r$, $\mathsf{th}{:}R_\theta$, $\mathsf{ph}{:}R_\varphi$, параметры робастности $\gamma = 0.3$, $\varepsilon = 0.01$.



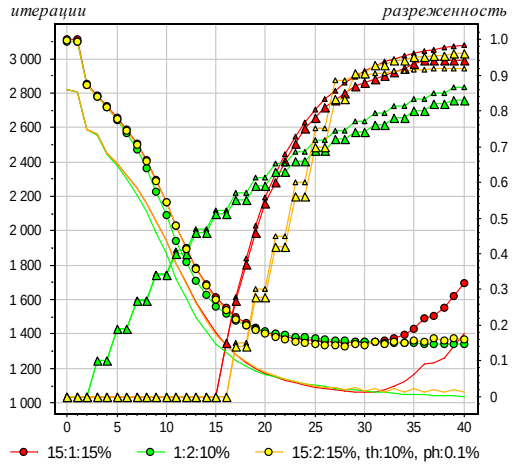
RuDis, PR, разреживание через 2 итерации



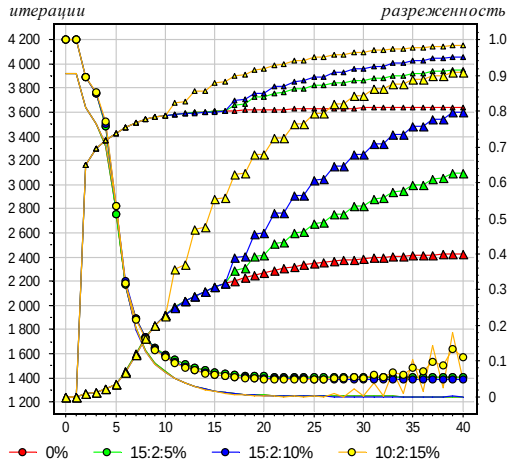
NIPS, PR, разреживание через 2 итерации



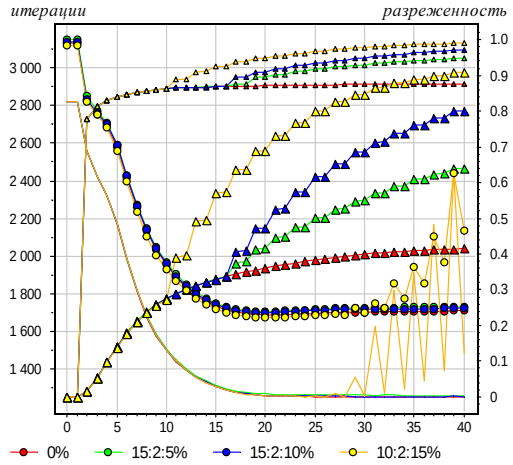
RuDis, PR, агрессивное разреживание



NIPS, PR, агрессивное разреживание



RuDis, SR, разреживание через 2 итерации



NIPS, SR, разреживание через 2 итерации

Рис. 6: Зависимость перплексии (○) и разреженности матриц Φ (▵) и Θ (△) от числа итераций при разреживании $p(t \mid d)$, $p(w \mid t)$. Параметры разреживания $i_0{:}\delta{:}r$, $\mathsf{th}{:}R_\theta$, $\mathsf{ph}{:}R_\varphi$, параметры робастности $\gamma = 0.01$, $\varepsilon = 0.01$.

При агрессивном разреживании или при использовании сэмплирования возможно разреживание распределений φ_t до 99%. При числе тем $|T| = 100$ это означает, что каждый термин в среднем относится только к одной теме.

В робастных алгоритмах с недостаточным априорным уровнем шума $\gamma = 0.01$ агрессивное разреживание может приводить к расходимости EM-алгоритма, рис. 6. Тонкие кривые без точек, проходящие чуть ниже кривых контрольной перплексии — это перплексия на обучающей выборке. Они показывают, что расходимость возникает синхронно на контроле и обучении, причём на обучении расходимость даже более заметна. Этот результат указывает на проблему выбора априорного уровня шума в робастных алгоритмах. В предлагаемой скорректированной модели PLSA такой проблемы не возникает, так как доля шума определяется автоматически и соответствует выбранным параметрам разреживания.

Сформулируем основные выводы:

1. Простой и вычислительно эффективный способ постепенного разреживания распределений позволяет достигать до 96% нулей без потери качества.

2. Неотъемлемой частью разреженной модели являются робастные компоненты, описывающие нетематические слова. В совмещении с разреживанием упрощенная робастная модель не уступает по качеству робастной модели с шумом и фоном, при этом является более эффективной с вычислительной точки зрения.

3. Обнуления должны начинаться не сразу, а только после сходимости модели, и происходить постепенно от итерации к итерации. В противном случае, разреживание может привести к разрушению модели. Заметим, что та же логика работает и в методе OBD: сначала строится полносвязная сеть, а затем небольшими порциями обнуляются синаптических веса, причем между обнулениями производится несколько дополнительных итераций, позволяющих модели сойтись к локальному оптимуму.

5 Аддитивная регуляризация тематических моделей

Задача адекватного описания текстов на естественном языке и выделения интерпретируемых тем, соответсвующих предметным областям, накладывает на вероятностную тематическую модель большое число требований. В предыдущих разделах были рассмотрены лишь некоторые из таких требований и приведены частные решения. Для построения гибких моделей, способных учитываеть все требования одновременно, необходим более общий подход.

5.1 EM-алгоритм для построения регуляризованной модели

В качестве такого подхода предлагается использовать *аддитивную регуляризацию тематических моделей* [9, 8]. Важной проблемой стандартной задачи построения тематической модели является неединственной и неустойчивость решения. Правдоподобие (2) зависит только от произведения $\Phi\Theta$, которое определено с точностью до линейного преобразования: $\Phi\Theta = (\Phi S)(S^{-1}\Theta)$, при условии, что матрицы $\Phi' = \Phi S$ и $\Theta' = S^{-1}\Theta$ также стохастические. Выбор преобразования S в EM-подобных алгоритмах никак не контролируется и зависит от случайного начального приближения.

В рамках данного подхода для обеспечения единственности и устойчивости решения используются проблемно-ориентированные регуляризаторы, формализующие адекватные предположения о матрицах Φ, Θ. В качестве основы берется модель PLSA, которая свободна от регуляризаторов и поэтому является удобной базой для построения сложных регуляризованных моделей.

Допустим, что наряду с правдоподобием (2) требуется максимизировать ещё n критериев $R_i(\Phi, \Theta),\ i = 1, \ldots, k$, называемых *регуляризаторами* [37]. Для решения задачи многокритериальной оптимизации будем максимизировать линейную комбинацию критериев $L(\Phi, \Theta)$ и $R_i(\Phi, \Theta)$ с неотрицательными *коэффициентами регуляризации* τ_i:

$$R(\Phi, \Theta) = \sum_{i=1}^{k} \tau_i R_i(\Phi, \Theta), \qquad L(\Phi, \Theta) + R(\Phi, \Theta) \ \to \ \max_{\Phi,\Theta}. \tag{27}$$

Коэффициенты регуляризации устанавливают баланс между оптимизируемыми критериями. Данная задача по-прежнему решается с помощью EM-алгоритма, но

вместо (5), (6) используется модифицированная формула М-шага:

$$\varphi_{wt} \propto \left(n_{wt} + \varphi_{wt} \frac{\partial R}{\partial \varphi_{wt}} \right)_+, \qquad \theta_{td} \propto \left(n_{dt} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right)_+, \tag{28}$$

где $(z)_+ = \max\{z, 0\}$ — операция положительной срезки.

В [9] показано, что система уравнений (4), (28), определяет стационарную точку задачи (27), (3).

5.2 Регуляризаторы разреживания, сглаживания, декоррелирования и сокращения незначимых тем

Рассмотрим регуляризаторы, направленные на учет лингвистических особенностей текста и повышение интерпретируемости тем.

Регуляризатор разреживания. Введем регуляризатор, формализующий гипотезу о том, что каждый документ и каждый термин связан с небольшим числом тем. Чем более разрежено распределение, тем меньше его энтропия. Максимальной энтропией обладает равномерное распределение. Поэтому будем максимизировать KL-дивергенцию между распределениями φ_t, θ_d и равномерными распределениями $\tilde{\beta} = (\tilde{\beta}_w)_{w \in W} = \left(\frac{1}{|W|}\right)_{w \in W}$, $\tilde{\alpha} = (\tilde{\alpha}_t)_{t \in T} = \left(\frac{1}{|T|}\right)_{t \in T}$ соответственно:

$$\sum_{t \in T} \mathrm{KL}_w(\tilde{\beta}_w \| \varphi_{wt}) \to \max_{\Phi}, \qquad \sum_{d \in D} \mathrm{KL}_t(\tilde{\alpha}_t \| \theta_{td}) \to \max_{\Theta},$$

где $\mathrm{KL}_i(p_i \| q_i) = \sum_{i \in I} p_i \log \frac{p_i}{q_i}$ — дивергенция Кульбака–Лейблера между дискретными распределениями $(p_i)_{i \in I}$ и $(q_i)_{i \in I}$; обозначение KL_i указывает на индекс суммирования i. Складывая два функционала с коэффициентами β_0, α_0 и удаляя из суммы константы, получим регуляризатор

$$R(\Phi, \Theta) = -\beta_0 \sum_{t \in T} \sum_{w \in W} \tilde{\beta}_w \ln \varphi_{wt} - \alpha_0 \sum_{d \in D} \sum_{t \in T} \tilde{\alpha}_t \ln \theta_{td} \to \max.$$

Обозначим $\alpha_t = \alpha_0 \tilde{\alpha}_t$, $\beta_w = \beta_0 \tilde{\beta}_w$. Формулы М-шага, согласно (28), имеют вид:

$$\varphi_{wt} \propto \left(n_{wt} - \beta_w \right)_+, \qquad \theta_{td} \propto \left(n_{dt} - \alpha_t \right)_+.$$

Идея энтропийной регуляризации была предложена в динамической тематической модели PLSA для разреживания распределений тем во времени при обработке

видеопотоков [38]. Однако возможность применения этой же техники для разреживания распределений φ_t и θ_d осталась незамеченной.

Указанная модификация M-шага приводит к тому, что на каждой итерации EM-алгоритма наименьшие значения в распределениях φ_t, θ_d обнуляются, при этом все остальные значения уменьшаются на определенную величину. Такой алгоритм очень близок к рассмотренным ранее стратегиям на основе метода OBD и отличается лишь постепенным приближением необнуляемых значений к нулю. Однако согласно экспериментам, эта деталь не влияет не результат.

Распределения $\tilde{\alpha}$, $\tilde{\beta}$ не обязательно задавать равномерными. В частности, в качестве распределения $\tilde{\beta}$ можно взять частоты слов в данной коллекции $\tilde{\beta}_w = n_w/n$, либо частоты слов в большой коллекции литературных текстов. Такой регуляризатор будет нацелен на удаление из тем частых слов коллекции или языка, не являющихся специфическими для данной темы.

Регуляризатор декоррелирования тем. Считается, что повышение различности тем улучшает интерпретируемость модели [34]. Действительно, каждая предметная область имеют свою терминологию, которая мало пересекается с терминологией других областей. Темы, которые близки по составу терминов, либо дублируют друг друга и посвящены одной и той же области, либо являются смесями нескольких областей. Оба случая нежелательны для модели.

Регуляризатор, минимизирующий ковариации между вектор-столбцами φ_t, φ_s,

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

приводит к формуле M-шага

$$\varphi_{wt} \propto \Big(n_{wt} - \tau \varphi_{wt} \sum_{s \in T \setminus t} \varphi_{ws} \Big)_+ .$$

Согласно этой формуле, вероятности φ_{wt} наиболее значимых тем слова 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 и удаляет из тем слова, часто использующиеся в других темах.

Регуляризатр сглаживания, введение фоновых тем. Модель с разреженными, сильно различными темами предназначена для описания предметных терминов в текстах коллекции, однако не предполагает описания общеупотребительных, нетематических слов. Поэтому предлагается ввести дополнительные *фоновые* темы B. Они призваны компенсировать действие разреживающего и декоррелирующего регуляризаторов на основные *предметные* темы S и взять на себя нетематические слова коллекции.

Фоновые темы должны содержать с ненулевой вероятностью все слова словаря, и особенно частые слова коллекции или языка. Кроме того, они должны присутствовать во всех словах документов. Поэтому введем для них регуляризатор сглаживания, обратный регуляризатору разреживания. Будем минимизировать KL-дивергенцию между фоновыми компонентами распределений φ_t, θ_d и распределениями $\tilde{\alpha}$, $\tilde{\beta}$:

$$R(\Phi, \Theta) = \beta_0 \sum_{t \in B} \sum_{w \in W} \tilde{\beta}_w \ln \varphi_{wt} + \alpha_0 \sum_{d \in D} \sum_{t \in B} \tilde{\alpha}_t \ln \theta_{td} \to \max .$$

Формулы M-шага для компонент распределений, соответствующих фоновым темам, согласно (28), принимают вид:

$$\varphi_{wt} \propto n_{wt} + \beta_w, \qquad \theta_{td} \propto n_{dt} + \alpha_t.$$

Те же сглаженные формулы оценки параметров используются в алгоритме сэмплирования Гиббса обучения модели LDA [33], однако выводятся из совершенно других предположений. При подходе аддитивной регуляризации сглаживание и разреживание описываются одинаково, если не вводить ограничений на знаки параметров β_w, α_t.

Тематическую модель с набором фоновых тем B для описания нетематических слов коллекции можно считать обобщением рассмотренных ранее робастных моделей, где в этих же целях используются дополнительные компоненты, никак не связанные с тематической моделью.

Регуляризатор сокращения незначимых тем. При построении модели могут появляться темы, к которым отнесено слишком мало слов. Такие темы не имеют интерпретации и являются избыточными. Чтобы гибко выявлять и исключать их

из модели, вводится регуляризатор, разреживающий распределение тем во всей коллекции $p(t) = \sum_d \theta_{td}\, p(d)$ и максимизируется дивергенция Кульбака-Лейблера между $p(t)$ и равномерным распределением. Формула регуляризованного М-шага в этом случае принимает вид

$$\theta_{td} \propto \left(n_{dt} - \tau\theta_{td}\frac{n_d}{n_t}\right)_+$$

где τ – коэффициент регуляризации. Согласно этой формуле, если число слов n_t, отнесенных к теме t во всей коллекции, мало, то вероятности этой темы понижаются для всех документов, вплоть до обнуления t-й строки матрицы Θ. Данный регуляризатор позволяет оптимизировать число тем, если начинать итерации с заведомо избыточного числа тем.

5.3 Многокритериальная оптимизация и оценивание качества

Траектории регуляризации. Для построения модели, удовлетворяющей большому числу требований, необходимо одновременное использование нескольких регуляризаторов. При линейном комбинировании регуляризаторов R_i возникает проблема выбора вектора коэффициентов $\tau = (\tau_i)_{i=1}^k$. Аналогичная проблема эффективно решается в эластичных сетях (elastic net) при комбинировании l_1- и l_2-регуляризации для задач регрессии и классификации [17]. В задачах тематического моделирования разнообразие регуляризаторов гораздо больше, и они могут влиять друг на друга. Поэтому стоит нетривиальная задача подбора траектории в пространстве коэффициентов регуляризации экспериментальным путём и формирования набора рекомендаций, позволяющих эффективно использовать аппарат аддитивной регуляризации.

Оценивание качества. Многокритериальный подход необходим не только на этапе построения тематической модели, но и на этапе оценивания качества. В экспериментах данного раздела точность описания коллекции документов по-прежнему контролировалась с помощью перплексии. Разреженность искомых распределений для моделей, разделяющих множество тем T на предметные S и фоновые B, оценивалась только по частям матриц Φ, Θ, соответствующих предметным темам.

Помимо этого, измерялась *интерпретируемость* тематической модели рядом показателей, не требующих работы экспертов. Во-первых, когерентностью, кото-

рая, как известно, хорошо коррелирует с человеческими оценками интерпретируемости [30, 31, 29]. Тема называется *когерентной*, если термины, наиболее частые в данной теме, неслучайно часто совместно встречаются рядом в документах коллекции. Во-вторых, был предложен ряд новых автоматически вычисляемых мер, основанных на предположении, что интерпретируемая тема имеет *ядро* из характерных слов, отличающих данную тему от остальных. Предполагается, что интерпретируемость темы тем лучше, чем больше терминов содержится в ядре, чем больше их суммарная вероятность (чистота темы) и чем больше вероятность встретить термины ядра именно в данной теме (контрастность темы). Введём эти показатели более формально.

Когерентность темы t измерялась как средняя *поточечная взаимная информация* по всем парам k наиболее вероятных слов темы t:

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

где w_i — i-й термин в порядке убывания φ_{wt}. Поточечная взаимная информация $\mathrm{PMI}(w_i, w_j)$ оценивает совместную встречаемость двух терминов:

$$\mathrm{PMI}(w_i, w_j) = \log \frac{p(w_i, w_j)}{p(w_i)p(w_j)},$$

где вероятность $p(w_i, w_j)$ пропорциональна числу раз, когда слова w_i, w_j встречаются в одном и том же документе.

Число k в большинстве работ полагают равным 10. Интересно оценить когерентность более глубоко, поэтому мы вычисляли ещё две оценки когерентности модели: при $k = 100$ и по ядрам тем. Когерентность может оцениваться как по самой коллекции D [29], так и по сторонней коллекции, например, по Википедии [32]. В наших экспериментах использовалась та же коллекция. Когерентность тематической модели оценивалась как средняя PMI_t по всем предметным темам $t \in S$.

Определим *ядро* W_t темы t как множество слов, для которых $p(t \mid w) > \kappa = 0.25$. Это множество терминов, которые имеют высокую вероятность в данной теме и низкие вероятности в остальных темах. На основе понятия ядра введем три показателя качества темы t:

- размер ядра: $|W_t|$;

- чистота темы: $\sum_{w \in W_t} p(w \mid t)$;
- контрастность темы: $|W_t|^{-1} \sum_{w \in W_t} p(t \mid w)$.

Контрастность показывает, насколько хорошо ядро темы отличает её от остальных тем. *Чистота* темы показывает, насколько велика доля терминов ядра внутри самой темы. Соответствующие показатели качества модели (размер ядра, чистота и контрастность) определим как средние по всем предметным темам $t \in S$.

5.4 Эксперименты

Применим подход аддитивной регуляризации на практике и построим тематическую модель с оптимальной комбинацией регуляризаторов сглаживания, разреживания, декоррелирования и сокращения незначимых тем. Нашей целью будет построение сильно разреженной модели при одновременном улучшении интерпретируемости тем и выделении общеупотребительных слов. Таким образом, требуется значительно улучшить несколько показателей качества, не ухудшив (или почти не ухудшив) правдоподобие модели.

Условия проведения экспериментов. Эксперименты проводились на коллекции англоязычных статей NIPS. Число итераций фиксировано равным 40, число тем $|T| = 100$, из них фоновых тем $|B| = 10$.

В экспериментах сравниваются стандартная модель PLSA и модели с различным сочетанием регуляризаторов. Результаты представляются в виде графиков зависимости различных показателей качества модели от номера итерации. На каждом рисунке сравниваются результаты двух моделей (серые и черные линии). Показатели качества выведены на трёх графиках друг под другом, с синхронизированными горизонтальными осями. Верхний график: по левой оси перплексия, по правой разреженности матриц Φ, Θ. Средний график: по левой оси размер ядра, по правой контрастность и чистота. Нижний график: когерентности ядра, top-10 и top-100.

ARTM позволяет комбинировать регуляризаторы в любых сочетаниях. Поэтому появляется возможность исследовать, как и на какие показатели качества влияет тот или иной регуляризатор.

Результаты. На рис. 7 сравниваются базовая модель PLSA и модель с декоррелированными предметными темами и сглаженными фоновыми. Сглаживающий регуляризатор здесь и далее использует равномерное распределение и параметры $\alpha = 0.8$ для профилей документов, $\beta = 0.1$ для профилей тем. Декоррелирование включается с первой итерации, коэффициент регуляризации подобран наибольшим, при котором ещё не происходит существенного увеличения перплексии. В результате втрое увеличивается чистота тем, в полтора раза когерентности top-10 и top-100. Перплексия по-прежнему сходится, но к чуть большему значению. Разреженность Θ явно не достаточна (55%), матрица Φ вообще не разрежена.

Сильной разреженности достигает модель с равномерным разреживающим регуляризатором для предметных тем и сглаживающим для фоновых (рис. 8). В отличие от декоррелирования, разреживание необходимо включать плавно. В данном эксперименте, начиная с 10-й итерации, коэффициент регуляризации постепенно изменяется так, чтобы на каждой итерации обнулялось 8% ненулевых значений в каждом векторе θ_d и 10% в каждом φ_t. Более раннее или более резкое разреживание может сильно ухудшать перплексию. При данной стратегии удается достичь разреженности Φ 96% и Θ 87% при несущественном ухудшении перплексии. На четверть улучшаются чистота и контрастность тем, когерентность ядра. При этом почти вдвое сокращается размер ядра. Это связано с тем, что примерно с 30-й итерации разреживание матрицы Φ переходит барьер 90% и начинает затрагивать наименее значимые слова из ядер тем.

На рис. 9 сравниваются два вида разреживания предметных тем в сглаженно-разреженной модели. Достигаемый уровень разреженности матриц Φ и Θ совпадает, однако показатели интерпретируемости ведут себя по-разному. Разреживание по равномерному распределению исключает из тем редкие шумовые термины и улучшает контрастность. Разреживание по распределению слов в коллекции улучшает все когерентности, а также почти втрое улучшает чистоту, т. е. суммарную вероятность терминов ядра в теме. Это происходит из-за очищения тем от общеупотребительных слов, которые за счет своего частого использования имеют высокие вероятности в теме, но не являются тематическими и не входят в ядро, т. к. употребляются в рамках многих тем.

Для одновременного достижения наиболее высоких показателей разреженности и интерпретируемости целесообразно комбинировать регуляризаторы разреживания и декоррелирования. Совмещение декоррелирвоания с равномерным разреживанием, представлено на рис. 9 и приводит к наилучшим результатам по совокупности критериев качества: перплексия меняется не сильно, разреженность достигает 96% для Φ и 87% для Θ; чистота 73%; контрастность 56%; когерентности top-10: 1.48, top100: 1.1, ядро: 1.28, средний размер ядра около 80. Совмещение декоррелирования с разреживанием по фону (рис. 11), уступает предыдущей модели по контрастности (52%), зато достигает существенно более высокой чистоты: 89%.

На рис. 12 демонстрируется действие регуляризатора постепенного сокращения числа тем в модели. Этот регуляризатор, так же как и разреживающий, лучше включать постепенно, на фоне сходимости итерационного процесса. В данном эксперименте модель начинает строиться при 100 темах, при этом начиная с 20-ой итерации, на каждой итерации удаляется 4% наименее значимых тем. За 100 итераций число тем сокращается до 20. Текущее число тем отображается на втором графике. Дополнительно выводится доля фона: доля слововхождений в коллекции, отнесенных к фоновым темам. Эта информация в совокупности с остальными критериями позволяет следить за тем, чтобы тематическая модель не выродилась в модель фона. Удаление наименее значимых тем и одновременный учет тех же самых тем для декоррелирования может приводить к несогласованности работы регуляризаторов и ухудшению перплексии модели. Поэтому итерации, на которых учитываются поправки регуляризаторов декоррелирования и сокращения незначимых тем, чередуются. Эксперимент показывает, что для коллекции NIPS минимально необходимое число тем равно 60: до этого значения сокращения тем происходят без потери перплексии, дальнейшие сокращения приводит к увеличению перплексии, а затем к вырождению модели.

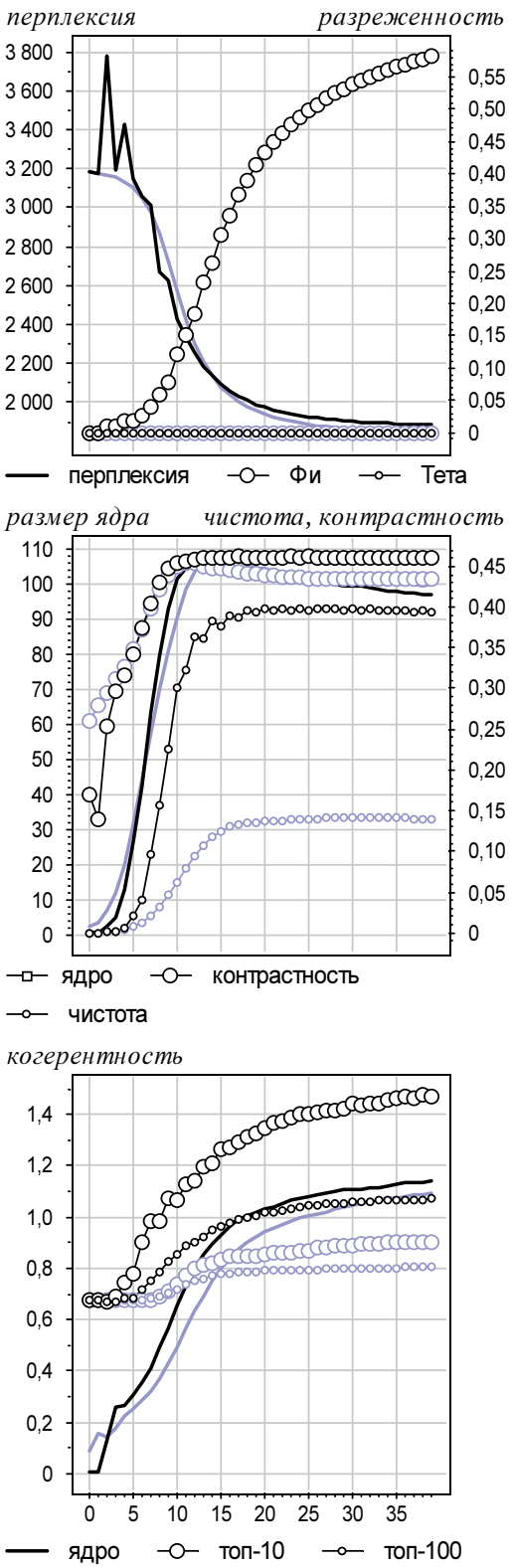


Рис. 7: Серый: PLSA. Чёрный: декоррелирование, сглаживание.

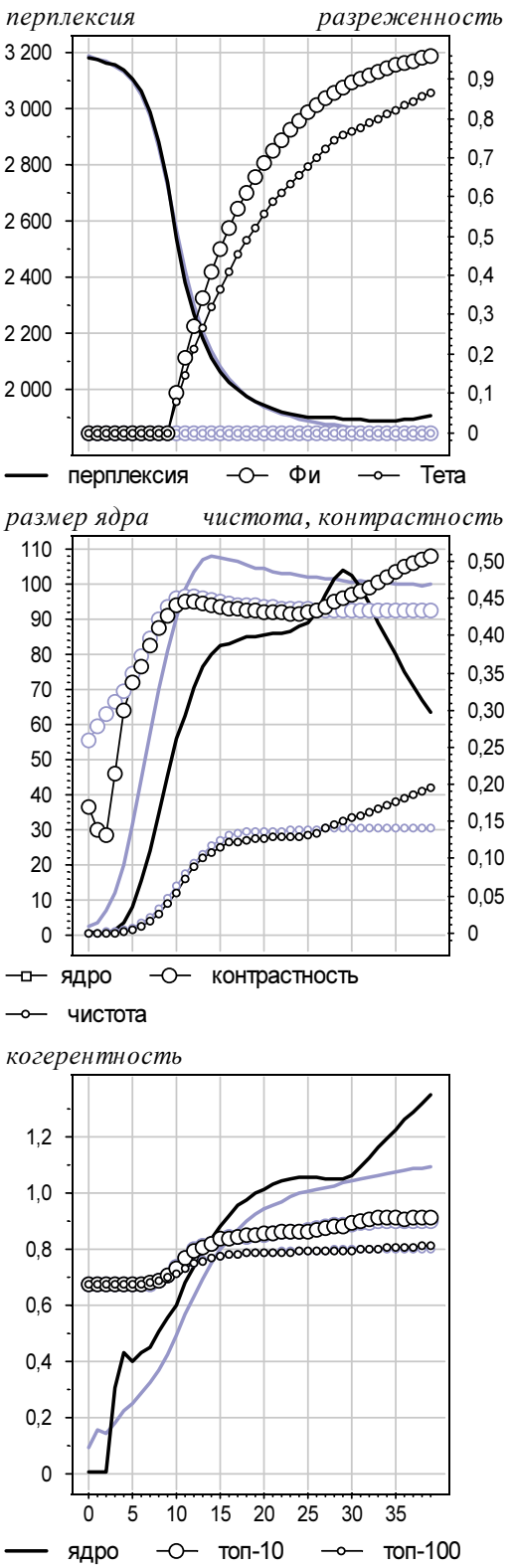


Рис. 8: Серый: PLSA. Чёрный: равномерное разреживание, сглаживание.

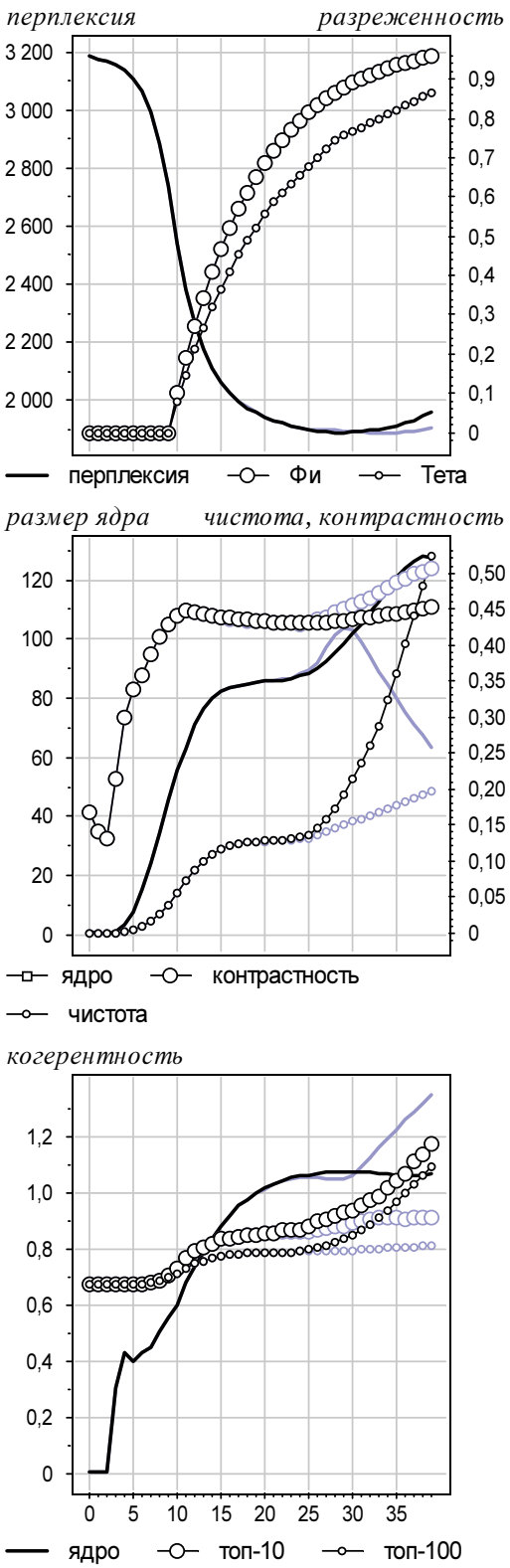


Рис. 9: Серый: равномерное разреживание. Чёрный: разреживание по фону.

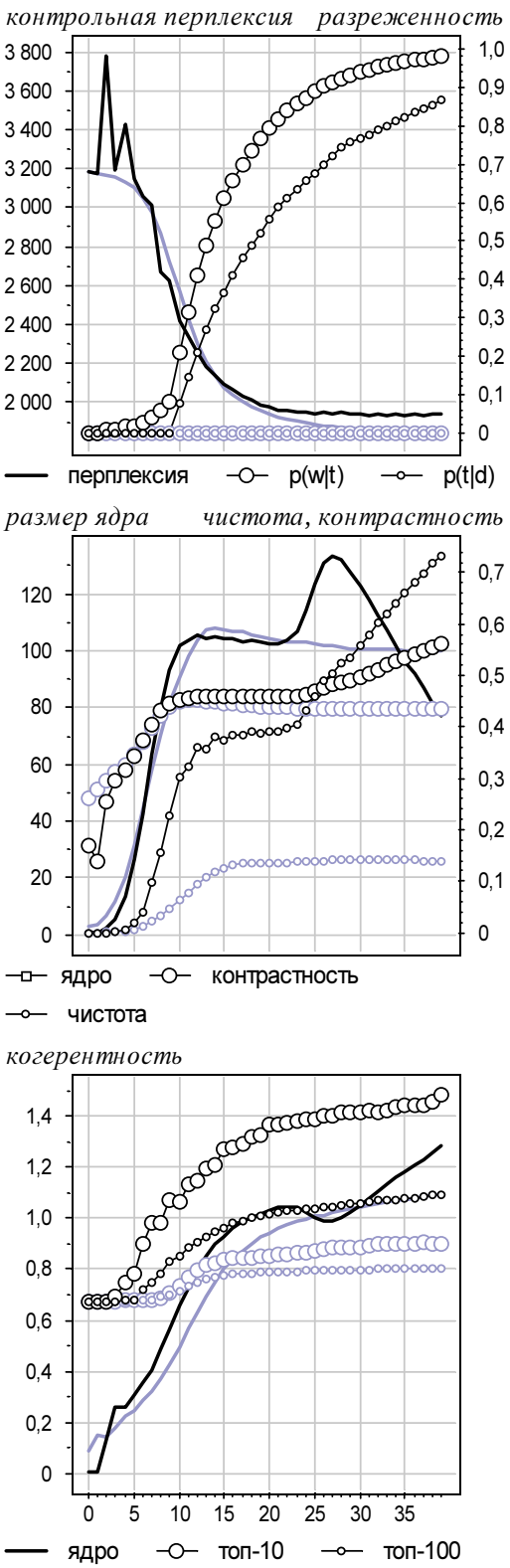


Рис. 10: Серый: PLSA. Чёрный: равномерное разреживание, сглаживание, декоррелирование.

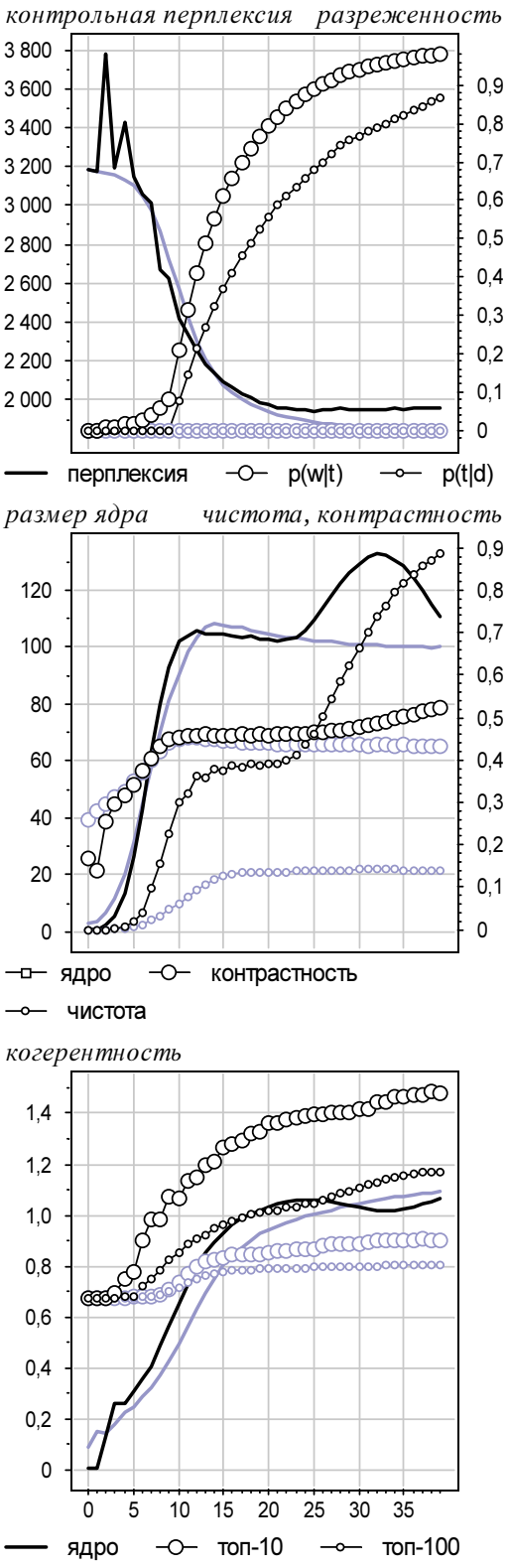


Рис. 11: Серый: PLSA. Чёрный: разреживание по фону, сглаживание, декоррелирование.

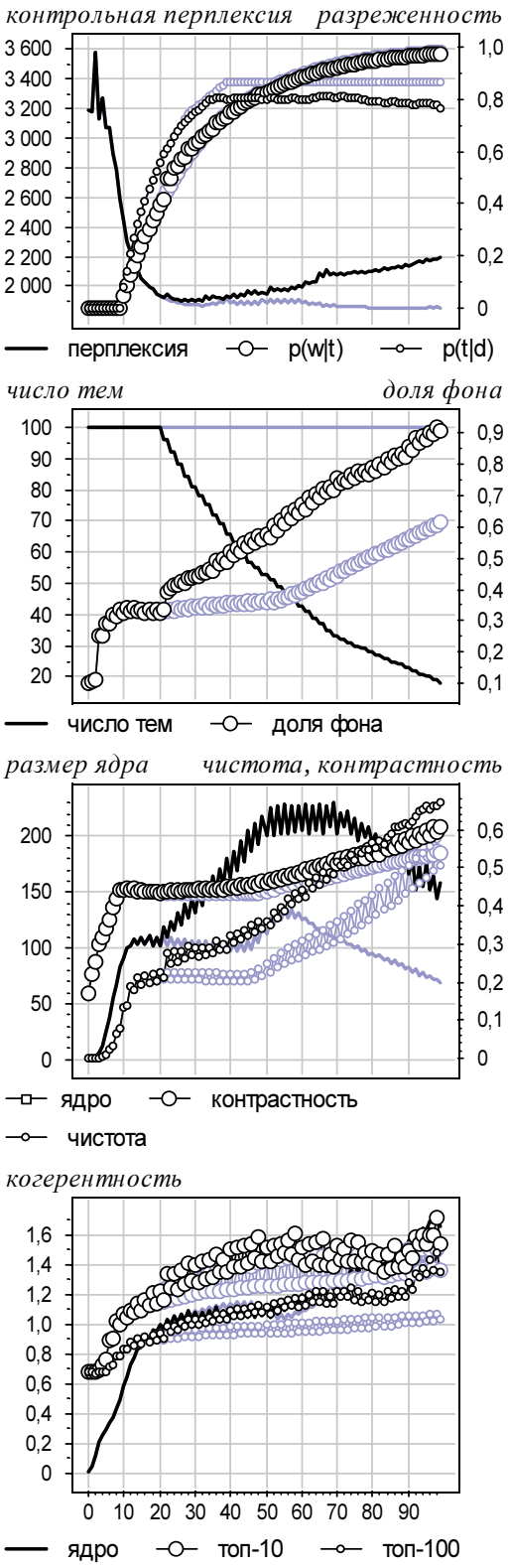


Рис. 12: Серый: разреживание, сглаживание, декоррелирование. Чёрный: разреживание, сглаживание, декоррелирование, сокращение тем.

Итак, построена регуляризованная тематическая модель, которая выделяет разреженные декоррелированные предметные темы и сглаженные фоновые. Эта модель существенно улучшает все измеряемые показатели интерпретируемости тем: чистоту, контрастность и когерентность. Интересно проанализировать подробнее словарный состав тем. В таблицах 4-5 представлены по 20 наиболее вероятных слов в некоторых темах базовой модели PLSA и предлагаемой регуляризованной модели. Слова упорядочены по убыванию вероятностей n_{wt}. Тематические термины, вошедшие в ядро темы, выделены жирным. В данной коллекции темы интерпретируются как задачи, подходы, методы, отвечающие тематике научной конференции NIPS.

Первое наблюдение состоит в том, что выделенные слова (ядра тем) являются определяющими для темы, в то время как по остальным словам понять тему гораздо труднее. Например, тема 50 посвящена распознаванию лиц. Для модели PLSA жирным выделены все слова с корнем face (лицо). Остальные слова – recognition (распознавание), representation (представление), figure (рисунок), model (модель) и т.д. – могут относиться ко многим другим темам NIPS.

Сравнение PLSA и регуляризованной модели показывает, что в ARTM слова с корнем face получают наибольшие вероятности. В топе появляется гораздо больше выделенных слов, и они, действительно, характерны для данной темы: Cottrell, Pentland (фамилии двух ученых, занимающихся распознаванием лиц), gesture (жестикуляция), lane (морщина), emotion (эмоции), и т.д. Аналогичные выводы можно сделать по теме 32, посвященной задаче ранжированию документов по запросу, и по большинству других тем.

Модели строились из одного и того же начального приближения, поэтому в большинстве случаев темы с одинаковыми номерами явно похожи друг на друга. Это позволяет избежать обычных сложностей сопоставления тем в двух моделях (например, с помощью венгерского алгоритма). Не совсем так происходит с темой 2. В модели ARTM она посвящена обработке музыкальных сигналов. В модели PLSA тема состоит из слов model (модель), prediction (предсказание), series (серия), neural (нейронный), data (данные) и других слов, нейтральных для коллекции NIPS и не позволяющих интерпретировать тему. Единственная тема в PLSA, содержащая в топе слова music (музыка) или melody (мелодия) – это тема 55. Тем не менее, она

Таблица 4: Сравнение тем в моделях PLSA и ARTM.

PLSA, тема 50	ARTM, тема 50	PLSA, тема 32	ARTM, тема 32
face	**face**	**query**	**mlp**
images	**faces**	set	**query**
faces	**facial**	**queries**	**queries**
recognition	**cottrell**	data	**cart**
set	**pentland**	algorithm	**documents**
image	**gesture**	learning	retrieval
based	**lane**	**documents**	relevant
hme	**emotion**	number	**document**
facial	**person**	performance	rank
representation	**steering**	words	sampling
view	**appearance**	mlp	instances
figure	**baluja**	**cart**	**splits**
model	**setpoint**	values	collection
experts	camera	cluster	gibbs
network	tracking	experiments	**sex**
human	**pose**	results	**ranking**
expert	**pomerleau**	relevant	ordering
space	**mouth**	**retrieval**	recursive
examples	**darrell**	classification	text
system	**lighting**	algorithms	axis

снова содержит много общеупотребительных слов коллекции, затрудняющих интерпретацию. Соответствующая тема в модели ARTM свободна от таких слов и целиком состоит из предметных терминов, относящихся к анализу музыки.

В таблице 5 представлена одна из фоновых тем, выделенных моделью ARTM. Все фоновые темы содержат термины, широко употребляемые во всей коллекции NIPS. При этом в некоторых фоновых темах доминируют слова, имеющие отношение к классификации, в некоторых – к вероятностным моделям, в некоторых – к нейронным сетям, и т. д.

Таблица 5: Сравнение тем в моделях PLSA и ARTM.

PLSA, тема 2	PLSA, тема 55	ARTM, тема 2	ARTM, фон
model	**music**	**estimator**	model
prediction	rules	**music**	data
series	note	**musical**	models
neural	representation	**notes**	parameters
models	neural	mozer	noise
data	events	**melody**	neural
estimation	net	**composition**	mixture
time	set	**bach**	prediction
function	time	**chorales**	set
method	**musical**	**melodic**	gaussian
nonlinear	figure	**jackknife**	likelihood
based	network	cooperative	networks
point	**notes**	**subnet**	test
points	input	**gem**	figure
estimator	**melody**	**melodies**	training
parameters	structure	**icl**	performance
error	harmony	**tonal**	network
algorithm	**tau**	**accent**	number
estimate	**pitch**	augmented	input
linear	temporal	piece	results

6 Заключение

В работе исследуются лингвистические свойства коллекций текстовых документов, и рассматриваются возможности их учета для построения интерпретируемых тематических моделей. Интерпретируемость является трудно формализуемым понятием и имеет множество аспектов. В данной работе предлагается формализация, основанная на понятии ядра темы – множества характерных слов, которые составляют терминологию соответствующей предметной области. Требование наличия выделенного ядра приводит к набору естественных предположений о свойствах тем: темы должны быть разреженными, существенно различными, свободными от общеупотребительных слов.

Для описания нетематических слов предложена опция робастности, которая заключается в дополнении тематической модели фоновой и шумовой компонентами. Данная модификация легко встраивается в любой EM-подобный алгоритм обучения модели. Эксперименты на русскоязычной и англоязычной коллекциях научных текстов показывают, что робастность существенно улучшает перплексию за счет альтернативного описания редких слов коллекции, не имеющих тематической окраски. При этом исчезает необходимость в использовании регуляризации Дирихле, главным образом, влияющей именно на редкие слова, сглаживая для них оценки.

Разреженность тематических моделей достигается с помощью простых и вычислительно эффективных стратегий постепенного обнуления наименее значимых элементов в распределениях Φ и Θ. Экспериментальное сравнение различных стратегий показывает, что обнуления важно начинать не сразу, а только после сходимости модели, и проводить постепенно от итерации к итерации. Данная методика имеет тесную аналогию с методом OBD постепенного обнуления синаптических связей в многослойных нейронных сетях. В сочетании с разреживанием предложена упрощенная робастная модель, в которой вес шумовой компоненты настраивается автоматически в результате разреживания тем. Данная модель не уступает по перплексии стандартным робастным моделям, при этом не требует дополнительных расходов по памяти и оказывается более эффективной с вычислительной точки зрения.

Исследования по робастным и разреженным моделям обобщаются с помощью подхода аддитивной регуляризации тематических моделей. Вводятся регуляризато-

ры разреживания, сглаживания, декоррелирования и сокращения незначимых тем модели. Их оптимальная комбинация приводит к гибкой тематической модели с неизбыточным числом тем, содержащей сильно различные предметные и сглаженные фоновые темы. Эксперименты показывают, что данная модель улучшает когерентность, чистоту, контрастность и разреженность тем без существенного ухудшения перплексии. Примеры наиболее вероятных слов в темах базовой и регуляризованной моделей говорят об очищении предметных тем от общеупотребительных слов коллекции. Специфика отдельных тем становится более понятной, улучшается их интерпретируемость. Нейтральные слова при этом концентрируются в фоновых темах. Таким образом, комбинирование регуляризаторов в вероятностных тематических моделях позволяет автоматически, без привлечения экспертов, учитывать лингвистические особенности данных и улучшать интерпретируемость тем.

Список литературы

[1] Воронцов К. В., Потапенко А. А. Робастные разреженные вероятностные тематические модели // Интеллектуализация обработки информации (ИОИ-2012): Докл.– Москва: Торус Пресс, 2012. С. 605-608.

[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] Потапенко А. А. Разреживание вероятностных тематических моделей // Математические методы распознавания образов: 16-ая Всеросс. конф.: Докл. М.: МАКС Пресс, 2013. С. 89.

[5] Потапенко А. А. Регуляризация вероятностной тематической модели для выделения ядер тем // Сборник тезисов XXI Международной научной конференции студентов, аспирантов и молодых ученых "Ломоносов-2014". Секция "Вычислительная математика и кибернетика". М.: МАКС Пресс, 2014.

[6] Воронцов К. В., Потапенко А. А. Многокритериальная регуляризация вероятностных тематических моделей для улучшения интерпретируемости тем и определения числа тем // Международная конференция по компьютерной лингвистике "Диалог". – 2014 (доклад принят).

[7] 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.

[8] Vorontsov K. V., Potapenko A. A. Tutorial on Probabilistic Topic Modeling: Additive Regularization for Stochastic Matrix Factorization // Analysis of Images, Social

Networks, and Texts AIST-2014. – Lecture Notes in Computer Science LNCS, Springer (to appear).

[9] Vorontsov K.V., Potapenko A. A. Additive Regularization of Topic Models // Machine Learning Journal, Special Issue "Data Analysis and Intelligent Optimization". Springer, 2014 (to appear).

[10] 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.

[11] Bahrani M, Sameti H. A New Bigram-PLSA Language Model for Speech Recognition // EURASIP Journal on Advances in Signal Processing, 2010.

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

[13] 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, 2006.– Vol. 19.– Pp. 241-248.

[14] Chen X., Qi Y., Bai B., Lin Q., Carbonell J. G. Sparse latent semantic analysis. // SDM.– 2011.– Pp. 474-485.

[15] 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.

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

[17] Friedman J. H., Hastie T., Tibshirani R. Regularization paths for generalized linear models via coordinate descent // Journal of Statistical Software.– 2010.– Vol. 33, no. 1.– Pp. 1-22.

[18] Griffiths T. L., Blei D. M., Steyvers M., Tenenbaum J. B. Integrating Topics and Syntax // Advances in Neural Information Processing Systems, 2005.

[19] Griffiths T. L., Steyvers M., Tenenbaum J. B. Topics in Semantic Representation // Psychological Review, vol. 114, 2007.

[20] Gruber A., Rosen-Zvi M., Weiss Y. Hidden Topic Markov Models // Artificial Intelligence and Statistics (AISTATS), San Juan, Puerto Rico, 2007.

[21] 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.

[22] Johnson M. PCFGs, topic models, adaptor grammars and learning topical collocations and the structure of proper names // Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, 2010. - Pp. 1148-1157.

[23] 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.

[24] Lau J. H., Baldwin T., Newman D. On Collocations and Topic Models // ACM Transactions on Speech and Language Processing (TSLP) – Special issue on multiword expressions: From theory to practice and use, part 2, vol. 10. – New York, NY, USA: ACM, 2013.

[25] LeCun Y., Denker J., Solla S., Howard R. E., Jackel L. D. Optimal brain damage // Advances in Neural Information Processing Systems II / Ed. by D. S. Touretzky.— San Mateo, CA: Morgan Kauffman, 1990. citeseer.ist.psu.edu/lecun90optimal.html.

[26] MacKay D., Linda C. Bauman Peto A Hierarchical Dirichlet Language Model // Natural Language Engineering, vol. 1, 1995. – Pp. 1-19.

[27] Masada T., Kiyasu S., Miyahara S. Comparing LDA with pLSI as a dimensionality reduction method in document clustering // Proceedings of the 3rd International Conference on Large-scale knowledge resources: construction and application.– LKR'08.– Springer-Verlag, 2008.– Pp. 13-26.

[28] McCallum A. K. Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering.— http://www.cs.cmu.edu/m̃ccallum/bow.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

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

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

[35] Than K., Ho T. B. Fully sparse topic models. // ECML/PKDD (1).– 2012.– Pp. 490-505.

[36] Teh Y. W., Newman D., Welling M. A collapsed variational bayesian inference algorithm for latent dirichlet allocation // NIPS.– 2006.– Pp. 1353-1360.

[37] Tikhonov A. N., Arsenin V. Y. Solution of ill-posed problems.– W. H. Winston, Washington, DC, 1977.

[38] 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.

[39] Wallach H. M. Topic modeling: beyond bag-of-words // Proceedings of the 23rd international conference on Machine learning, 2006. – Pp. 977-984.

[40] Wallach H., Mimno D., McCallum A. Rethinking LDA: Why priors matter // Advances in Neural Information Processing Systems 22 / Ed. by Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, A. Culotta.– 2009.– Pp. 1973-1981.

[41] Wang X., Mccallum A. A note on topical n-grams // Technical Report UM-CS-2005-071, University of Massachusetts, 2005.

[42] Wang X., Mccallum A, Wei X. Topical n-grams: Phrase and topic discovery, with an application to information retrieval // In Proceedings of the 7th IEEE International Conference on Data Mining, 2007.

[43] Wang Y. Distributed Gibbs sampling of latent dirichlet allocation: The gritty details.– 2008.

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

[45] Wang Q., Xu J., Li H., Craswell N. Regularized latent semantic indexing. // SIGIR.– 2011.– Pp. 685-694.

[46] Wilson A.T., Chew P.A. Term weighting schemes for Latent Dirichlet Allocation // 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. 465-473

[47] Wu Y., Ding Y., Wang X., Xu J. A comparative study of topic models for topic clustering of chinese web news // Computer Science and Information Technology (ICCSIT), 2010 3rd IEEE International Conference on.– Vol. 5.– july 2010.– Pp. 236-240.

[48] Zhu J., Xing E. P. Sparse topical coding.– 2012.