КОМПЬЮТЕРНЫЕ ИССЛЕДОВАНИЯ
И МОДЕЛИРОВАНИЕ 2012 Т. 4 № 4 С. 693–706

Ки&М

МАТЕМАТИЧЕСКИЕ ОСНОВЫ И ЧИСЛЕННЫЕ МЕТОДЫ МОДЕЛИРОВАНИЯ

УДК: 004.852

Регуляризация, робастность и разреженность вероятностных тематических моделей

К. В. Воронцов[1,a], А. А. Потапенко[2,b]

[1]Лаборатория «РУКОНТ-ФизТех» ФУПМ МФТИ,
Россия, 141700, г. Долгопрудный, Институтский переулок, д. 9
[2]ВМК МГУ,
Россия, 119991 ГСП-1 г. Москва, Ленинские горы, МГУ имени М. В. Ломоносова, факультет ВМК

E-mail: [a] voron@forecsys.ru, [b] anya_potapenko@mail.ru

Получено 06 сентября 2012 г.

Предлагается обобщенное семейство вероятностных тематических моделей коллекций текстовых документов, в котором эвристики регуляризации, сэмплирования, частого обновления параметров, робастности относительно шума и фона могут включаться независимо друг от друга в любых сочетаниях, порождая как известные модели PLSA, LDA, CVB0, SWB, так и новые. Показано, что робастная тематическая модель на основе PLSA, разделяющая термины на тематические, шумовые и фоновые, не нуждается в регуляризации и обеспечивает разреженность искомых дискретных распределений тем в документах и терминов в темах.

Ключевые слова: компьютерный анализ текстов, тематическое моделирование, вероятностный латентный семантический анализ, EM-алгоритм, латентное размещение Дирихле, сэмплирование Гиббса, байесовская регуляризация, перплексия, робастность

Regularization, robustness and sparsity of probabilistic topic models

K. V. Vorontsov[1], A. A. Potapenko[2]

[1]RUKONT-PhysTech Laboratory, CMAM department, MIPT, 9 Institutskii per., Dolgoprudny, Moscow Region, 141700, Russia
[2]CMC department, Moscow State University, Leninskie gory, Moscow, 119991, Russia

Abstract. — We propose a generalized probabilistic topic model of text corpora which can incorporate heuristics of Bayesian regularization, sampling, frequent parameters update, and robustness in any combinations. Well-known models PLSA, LDA, CVB0, SWB, and many others can be considered as special cases of the proposed broad family of models. We propose the robust PLSA model and show that it is more sparse and performs better that regularized models like LDA.

Keywords: text analysis, topic modeling, probabilistic latent semantic analysis, EM-algorithm, latent Dirichlet allocation, Gibbs sampling, Bayesian regularization, perplexity, robusteness

Citation: *Computer Research and Modeling*, 2012, vol. 4, no. 4, pp. 693–706 (Russian).

Работа выполнена при поддержке Министерства образования и науки РФ (Государственный контракт 07.524.11.4002) и Российского фонда фундаментальных исследований (проект № 11-07-00480).

Введение

Тематическое моделирование (topic modeling) — одно из приложений машинного обучения к анализу текстов, активно развивающееся с конца 90-х годов. *Тематическая модель* коллекции текстовых документов определяет, к каким темам относится каждый документ и какие слова (термины) образуют каждую тему. Число тем в большинстве приложений заранее неизвестно и является одним из важнейших параметров модели. Таким образом, тематическая модель — это результат совместной кластеризации документов и терминов по кластерам-темам.

Вероятностные тематические модели (ВТМ) осуществляют «мягкую» кластеризацию, позволяя документу или термину относиться сразу к нескольким темам с различными вероятностями. ВТМ описывает каждую тему дискретным распределением на множестве терминов, каждый документ — дискретным распределением на множестве тем. Предполагается, что коллекция документов — это последовательность терминов, выбранных случайно и независимо из смеси таких распределений, и ставится задача восстановления компонент смеси по выборке.

В информационном поиске документы принято представлять векторами, координаты которых соответствуют словам, а значения — статистическим характеристикам слов, например частотам или tf-idf. Поиск документов по коротким запросам реализуется путем поиска векторов, в которых часто встречаются слова запроса [Маннинг и др., 2011].

В тематическом моделировании документы представляются векторами тем. Как правило, число тем много меньше числа слов в документе, поэтому происходит сильное сжатие документа при сохранении наиболее существенной информации о его тематике. Это позволяет быстро искать документы схожей тематики, задавая в качестве запроса документ или длинный фрагмент текста. Векторами тем представляются также связанные с документами объекты: термины, авторы, научные группы, организации, конференции, журналы, сайты и т. д., что позволяет задавать в качестве запроса любой объект или совокупность объектов и искать по ним объекты того же или другого типа, имеющие схожую тематику.

В данной работе рассматриваются основные разновидности ВТМ, считающиеся классическими. Существуют и более сложные модели, в которых учитываются марковские зависимости в последовательностях терминов, связи между документами через авторство или ссылки, плавные изменения тематики во времени, иерархические отношения между темами и другие особенности текстовых коллекций. Многочисленные разновидности вероятностных тематических моделей описаны в обзоре [Daud et al., 2010].

Исходными данными для тематического моделирования является множество (*коллекция*) текстовых документов D и множество (*словарь*) терминов W. Каждый документ $d \in D$ представляется последовательностью терминов $(w_1, \ldots, w_{n_d})$ из W, где n_d — длина документа. Через n_{dw} обозначается число вхождений термина w в документ d.

Основные предположения вероятностных тематических моделей [Hofmann, 1999; Blei et al., 2003].

1. Предполагается, что существует конечное множество тем T и коллекция порождается дискретным распределением $p(d, w, t)$ на $D \times W \times T$. Переменные d и w являются наблюдаемыми, переменная t — латентной, т. е. появление каждой пары (d, w) связано с некоторой неизвестной темой t. Построить тематическую модель коллекции — означает найти множество тем T, условные распределения $p(w \mid t) \equiv \phi_{wt}$ для каждой темы $t \in T$ и $p(t \mid d) \equiv \theta_{td}$ для каждого документа $d \in D$.

2. Предполагается, что распределение вероятностей терминов $p(w \mid d, t)$ зависит только от темы t, но не от документа d (*гипотеза условной независимости*):

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

3. Предполагается, что для выявления тематики достаточно знать, какие термины встречаются в каких документах, но не важен ни порядок терминов в документах (*гипотеза «мешок слов»*), ни порядок документов в коллекции (*гипотеза «мешок документов»*). Другими словами, предполагается, что тематику документа можно узнать даже после случайной перестановки терминов, хотя для человека такой текст теряет смысл. Формально это означает, что каждый документ $d \in D$ рассматривается как выборка терминов, порождаемых случайно и независимо из распределения

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

Условная вероятность в левой части оценивается по коллекции как $f_{dw} = n_{dw}/n_d$, поэтому построение ВТМ можно рассматривать также как задачу поиска разложения матрицы $F = (f_{dw})_{D \times W}$ в произведение двух неотрицательных нормированных матриц меньшего размера $\Phi = (\phi_{wt})_{W \times T}$ и $\Theta = (\theta_{td})_{T \times D}$.

Для определения параметров модели Φ, Θ максимизируется правдоподобие

$$L(\Theta, \Phi) = \sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td} \;\to\; \max_{\Theta, \Phi} \tag{3}$$

при ограничениях неотрицательности $\theta_{td} \geqslant 0$, $\phi_{wt} \geqslant 0$ и нормировки

$$\sum_{w \in W} \phi_{wt} = 1, \quad \sum_{t \in T} \theta_{td} = 1.$$

Качество вероятностных тематических моделей принято оценивать с помощью *перплексии* контрольных данных, которая используется также в компьютерной лингвистике. Она определяется через правдоподобие контрольной коллекции D' документов, не включенных в обучающую коллекцию (чем меньше перплексия, тем лучше):

$$\mathcal{P}(D') = \exp\Big(-\frac{1}{n} \sum_{d \in D'} \sum_{w \in d''} n_{dw} \ln p(w \mid d)\Big), \tag{4}$$

Каждый контрольный документ $d \in D'$ случайным образом делится на две половины, d' и d''. Параметры ϕ_{wt} оцениваются по обучающей выборке D. Параметры θ_{td} оцениваются по d'. Перплексия вычисляется по вторым половинам d'' контрольных документов.

Численное значение перплексии интерпретируется следующим образом. Если термины w_i порождаются из равномерного распределения $p(w)$ на словаре мощности V, то перплексия такого текста сходится к V с ростом его длины. Чем сильнее распределение p отличается от равномерного, тем меньше перплексия. Чем сильнее модель p отличается от генерирующего распределения, тем больше перплексия. В нашем случае вместо $p(w)$ используются условные вероятности терминов $p(w \mid d)$ и интерпретация немного другая: если каждый документ генерируется из V равновероятных терминов (возможно, различных в разных документах), то перплексия сходится к V.

Для тематических моделей, как и для других моделей машинного обучения, важной характеристикой является способность к обобщению. Если перплексия контрольной коллекции существенно выше перплексии обучающей коллекции, то говорят, что модель переобучается. Адекватных теоретических оценок переобучения для ВТМ на данный момент не существует. Качество моделей принято оценивать и сравнивать эмпирически на общедоступных коллекциях документов.

Краткий обзор литературы и цели исследования

В [Hofmann, 1999] предложен метод решения оптимизационной задачи (3) на основе EM-алгоритма, названный *вероятностным латентным семантическим анализом* (probabilistic latent semantic analysis, PLSA). Эксперименты [Blei et al., 2003] показали, что PLSA склонен к переобучению и не позволяет естественным образом добавлять в модель новые документы.

Также в [Blei et al., 2003] предложена модель *латентного размещения Дирихле* (latent Dirichlet allocation, LDA), которая до сих пор считается основной для BTM. В LDA указанные недостатки устраняются с помощью байесовской регуляризации — предположения, что векторы параметров ϕ_t, θ_d порождаются априорными распределениями Дирихле. Известны сотни модификаций модели LDA, учитывающих различные специфические особенности текстовых коллекций.

Одним из распространенных методов оптимизации параметров LDA является *сэмплирование Гиббса* (Gibbs sampling, GS) [Steyvers, Griffiths, 2004]. В [Asuncion et al., 2009] показано, что другие известные методы отличаются в основном способом сглаживания частотных оценок вероятностей, причем все они становятся практически эквивалентными, если оптимизировать гиперпараметры априорных распределений, как предлагается в [Wallach et al., 2009].

В [Chemudugunta et al., 2006] предложена робастная модель с шумовой и фоновой компонентами (special words with background, SWB), предполагающая, что некоторая доля терминов в документе может объясняться не тематикой, а индивидуальными особенностями документа (шумом) или общими особенностями всей коллекции (фоном). Для настройки этой модели использовалась модификация LDA-GS.

В данной работе вводится обобщенное семейство EM-подобных методов, в котором опции регуляризации, сэмплирования, оптимизации гиперпараметров, робастности относительно шума и фона могут включаться независимо друг от друга в любых сочетаниях, порождая как стандартные, так и новые BTM.

Нетривиальным и неожиданным результатом сравнения этих моделей оказалось то, что робастные модели не нуждаются в регуляризации. Этот факт позволяет лучше понять роль регуляризации в BTM и получить ряд практических преимуществ.

1. Повышение обобщающей способности при регуляризации [Blei et al., 2003] является кажущимся и вызвано лишь увеличением правдоподобия редких и новых терминов на контрольных данных. Эксперименты показывают, что если в контроле новых терминов нет, то включение регуляризации слабо влияет на результат (рис. 1). Робастная модель более точно описывает правдоподобие таких терминов (рис. 3).

2. Байесовская регуляризация сглаживает частотные оценки вероятностей, в результате матрицы Φ и Θ не содержат нулевых значений. Это противоречит интуитивно очевидной *гипотезе разреженности*: документ, как правило, относится к небольшому числу тем; тема определяется относительно небольшим числом терминов. Робастный PLSA строит разреженные Φ и Θ и позволяет управлять степенью их разреженности.

3. Перенос опций частого обновления параметров и сэмплирования из LDA-GS в PLSA решает проблему добавления новых документов и заметно ускоряет его. В данной работе предлагается эвристика *экономного сэмплирования*, которая ускоряет его еще сильнее.

Робастность, разреженность и ускоренное сэмплирование необходимы для создания алгоритмов оптимизации BTM, масштабируемых для обработки очень больших коллекций.

Тематическая модель PLSA

Для решения задачи (3) применяется EM-алгоритм — итерационный процесс, в котором каждая итерация состоит из шагов E (expectation) и M (maximization) [Dempster et al., 1977].

Допустим, что перед первой итерацией задано начальное приближение Θ, Φ.

На E-шаге для каждой пары (d, w) по формуле Байеса вычисляются условные вероятности тем H_{dwt} при фиксированных параметрах Θ, Φ:

$$H_{dwt} \equiv p(t \mid d, w) = \frac{\phi_{wt}\theta_{td}}{\sum_s \phi_{ws}\theta_{sd}}. \tag{5}$$

На M-шаге вычисляется приближенное решение задачи максимизации правдоподобия при фиксированных H_{dwt}. Чтобы получить приближенное решение M-шага, запишем лагранжиан задачи (3) при ограничениях нормировки, проигнорировав ограничения неотрицательности (после убедимся, что решение действительно неотрицательно):

$$\mathcal{L}(\Theta, \Phi) = \sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{t\in T} \phi_{wt}\theta_{td} - \sum_{t\in T} \lambda_t\Big(\sum_{w\in W} \phi_{wt} - 1\Big) - \sum_{d\in D} \mu_d\Big(\sum_{t\in T} \theta_{td} - 1\Big).$$

Продифференцировав лагранжиан по ϕ_{wt} и приравняв нулю производную, получим

$$\lambda_t = \sum_{d\in D} n_{dw} \frac{\theta_{td}}{p(w \mid d)}. \tag{6}$$

Домножим обе части равенства (6) на ϕ_{wt}, просуммируем по всем $w \in W$, применим условие нормировки вероятностей ϕ_{wt} в левой части и выделим переменную H_{dwt} в правой части. Получим $\lambda_t = \sum\limits_{d\in D}\sum\limits_{w\in W} n_{dw}H_{dwt}$. Снова домножим (6) на ϕ_{wt}, выделим H_{dwt} в правой части и выразим ϕ_{wt} из левой части, подставив уже известное выражение для λ_t. Получим

$$\phi_{wt} = \frac{\sum\limits_{d\in D} n_{dw}H_{dwt}}{\sum\limits_{w'\in W}\sum\limits_{d\in D} n_{dw'}H_{dw't}}.$$

Обозначив числитель через $\hat{n}_{wt}$, перепишем это выражение в следующем виде:

$$\phi_{wt} = \frac{\hat{n}_{wt}}{\hat{n}_t}, \qquad \hat{n}_t = \sum_{w\in W} \hat{n}_{wt}, \qquad \hat{n}_{wt} = \sum_{d\in D} n_{dw}H_{dwt}. \tag{7}$$

Проделав аналогичные действия с производной лагранжиана по θ_{td}, получим

$$\theta_{td} = \frac{\hat{n}_{dt}}{\hat{n}_d}, \qquad \hat{n}_d = \sum_{t\in T} \hat{n}_{dt}, \qquad \hat{n}_{dt} = \sum_{w\in d} n_{dw}H_{dwt}. \tag{8}$$

Заметим, что если начальные приближения θ_{td} и ϕ_{wt} неотрицательны, то и после каждой итерации они будут оставаться неотрицательными, несмотря на то, что ограничение неотрицательности было проигнорировано при решении задачи.

Если рассматривать коллекцию документов как выборку *токенов* — троек $(d, w, t) \in D \times W \times T$, то $n_{dwt} = n_{dw}H_{dwt}$ есть оценка числа вхождений термина w в документе d, связанных с темой t. Соответственно, переменные $\hat{n}_*$ интерпретируются как *счетчики* числа токенов: $\hat{n}_{dt}$ — оценка числа токенов в документе d, связанных с темой t; $\hat{n}_{wt}$ — оценка числа токенов термина w, связанных с темой t; $\hat{n}_t$ — оценка общего числа токенов в коллекции, связанных с темой t; $\hat{n}_d = n_d$, поэтому его можно не вычислять на каждой итерации.

Хранения трехмерной матрицы H_{dwt} и передачи ее из E-шага в M-шаг можно избежать. Переменные $\hat{n}_{wt}$ и $\hat{n}_{dt}$ вычисляются на M-шаге в цикле по всем документам $d \in D$ и всем терминам $w \in d$. Внутри этого цикла переменные H_{dwt} можно вычислять по формуле E-шага в тот момент, когда они понадобятся. Таким образом, E-шаг встраивается внутрь M-шага без дополнительных вычислительных затрат. Именно этот вариант реализации показан в Алгоритме 1.

Алгоритм 1 ЕМ-алгоритм для тематической модели PLSA.

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

1: **повторять**
2: обнулить $\hat{n}_{wt}$, $\hat{n}_{dt}$, $\hat{n}_t$ для всех $d \in D$, $w \in W$, $t \in T$;
3: **для всех** $d \in D$, $w \in d$
4: $Z := \sum_{t \in T} \phi_{wt}\theta_{td}$;
5: **для всех** $t \in T$ таких, что $\phi_{wt}\theta_{td} > 0$
6: увеличить $\hat{n}_{wt}$, $\hat{n}_{dt}$, $\hat{n}_t$ на $\delta = n_{dw}\phi_{wt}\theta_{td}/Z$;
7: $\phi_{wt} := \hat{n}_{wt}/\hat{n}_t$ для всех $w \in W$, $t \in T$;
8: $\theta_{td} := \hat{n}_{dt}/n_d$ для всех $d \in D$, $t \in T$;
9: **пока** Θ и Φ не стабилизируются.

Частое обновление параметров

В ЕМ-алгоритме нет необходимости очень точно решать задачу максимизации правдоподобия на каждом М-шаге. Достаточно сместиться в направлении максимума и затем выполнить Е-шаг. Модификация ЕМ-алгоритма, при которой Е-шаг выполняется чаще, называется *обобщенным ЕМ-алгоритмом* (generalized EM-algorithm, GEM). Для него справедливы те же доказательства сходимости, что и для основного варианта ЕМ-алгоритма [Dempster et al., 1977].

Обобщенный ЕМ-алгоритм в случае PLSA сводится к более частому обновлению параметров θ_{td} и ϕ_{wt} по значениям счетчиков $\hat{n}_{wt}$ и $\hat{n}_{dt}$. В Алгоритме 1 это происходит после каждого просмотра всей коллекции. На больших коллекциях более частые обновления должны повышать скорость сходимости. Обновления можно делать после каждой пары (d, w) или после заданного числа пар (d, w) или после каждого документа. В Алгоритме 2 выбор условия обновления на шаге 9 оставлен на усмотрение разработчика.

Замечания к Алгоритму 2.

1. На первой итерации частые обновления не делаются, чтобы в счетчиках накопилась информация по всей коллекции. В противном случае оценки параметров θ_{td} и ϕ_{wt} по начальному фрагменту выборки могут оказаться хуже начального приближения.

2. Начиная со второй итерации, для каждой пары (d, w) из счетчиков $\hat{n}_{wt}$ и $\hat{n}_{dt}$ вычитается n_{dwt} — то самое значение δ, которое было к ним прибавлено при обработке пары (d, w) на предыдущей итерации. Таким образом, счетчики $\hat{n}_{wt}$ и $\hat{n}_{dt}$ всегда содержат актуальное значение, сформированное при последнем просмотре всей коллекции.

Сэмплирование

В Алгоритме 2 для каждой пары (d, w) хранится весь массив значений n_{dwt}, $t \in T$. Даже при небольшом числе тем такой расход памяти на хранение каждой пары (d, w) может оказаться неприемлемым. В то же время, согласно гипотезе разреженности, вхождение термина w в документ d связано, скорее всего, с небольшим числом тем. Эксперименты показывают, что тривиальное отбрасывание близких к нулю значений n_{dwt} на каждом шаге может приводить к накоплению большой систематической ошибки и смещению модели.

В таком случае лучше использовать сэмплирование — для каждой пары (d, w) генерировать s случайных тем t_{dwi}, $i = 1, \ldots, s$, из распределения $p(t \mid d, w) = H_{dtw} = \phi_{wt}\theta_{td}/Z$. Тогда число

Алгоритм 2 Обобщенный EM-алгоритм для тематической модели PLSA.

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

1: обнулить $\hat{n}_{wt}$, $\hat{n}_{dt}$, $\hat{n}_t$, $\hat{n}_d$, n_{dwt} для всех $d \in D$, $w \in W$, $t \in T$;
2: **повторять**
3: **для всех** $d \in D$, $w \in d$
4: $Z := \sum_{t \in T} \phi_{wt}\theta_{td}$;
5: **для всех** $t \in T$ таких, что $n_{dwt} > 0$ или $\phi_{wt}\theta_{td} > 0$
6: $\delta := n_{dw}\phi_{wt}\theta_{td}/Z$;
7: увеличить $\hat{n}_{wt}$, $\hat{n}_{dt}$, $\hat{n}_t$, $\hat{n}_d$ на $(\delta - n_{dwt})$;
8: $n_{dwt} := \delta$;
9: **если** не первая итерация и пора обновить параметры Φ, Θ **то**
10: $\phi_{wt} := \hat{n}_{wt}/\hat{n}_t$ для всех $w \in W$, $t \in T$ таких, что $\hat{n}_{wt}$ изменился;
11: $\theta_{td} := \hat{n}_{dt}/\hat{n}_d$ для всех $d \in D$, $t \in T$ таких, что $\hat{n}_{dt}$ изменился;
12: **пока** Θ и Φ не стабилизируются.

ненулевых значений n_{dwt} будет невелико, и в то же время оценки будут несмещенными. Сэмплирование можно рассматривать как замену условного распределения $p(t|d,w)$ его эмпирической оценкой по сгенерированной случайной выборке длины s:

$$\hat{p}(t|d,w) = \frac{1}{s}\sum_{i=1}^{s}[t_{dwi} = t]. \tag{9}$$

Сэмплирование в Алгоритме 2 реализуется путем трех небольших модификаций:

1) перед шагом 5 сэмплируется s тем $t = t_{dwi}$, $i = 1, \ldots, s$ из распределения $p(t\,|\,d,w)$;

2) на шаге 5 цикл по всем темам $t \in T$ заменяется циклом по $t = t_{dwi}$, $i = 1, \ldots, s$;

3) на шаге 6 вычисляется $\delta := n_{dw}/s$.

Таким образом, в обычном PLSA n_{dw} вхождений термина w в документ d распределяются между всеми $|T|$ темами пропорционально вероятностям $p(t\,|\,d,w)$, тогда как при сэмплировании задействуется не более s тем.

Сэмплирование Гиббса для модели LDA [Steyvers, Griffiths, 2004] во многом аналогично сэмплированию в модифицированном Алгоритме 2. Его строгий вывод приводится в [Wang, 2008]. PLSA-GEM с сэмплированием (модифицированный Алгоритм 2) и LDA-GS (Алгоритм 3) имеют несколько отличий, но только одно из них оказывается существенным с точки зрения качества модели.

1. В LDA-GS жестко фиксируется параметр $s = n_{dw}$. Однако *гипотеза разреженности* предполагает, что появление термина w в документе d вряд ли может быть связано с большим числом тем. В наших экспериментах $s = 5$ тем оказалось достаточно, но одной темы явно мало, см. рис. 2. Эвристика *экономного сэмплирования* повышает эффективность алгоритма как по скорости, так и по памяти, не ухудшая качество модели.

2. В LDA-GS параметры ϕ_{wt} и θ_{td} обновляются предельно часто — после обработки каждого вхождения термина w в документ d. Эксперименты показывают, что можно делать обновления чуть реже — после каждой пары (d,w), это не влияет на качество модели.

3. В LDA-GS перед сэмплированием на шаге 7 производится уменьшение счетчиков на единицу. Тем самым в оценке распределений не учитывается i-е вхождение термина w в документ d, для которого сэмплируется тема t_{dwi}. Эта особенность алгоритма следует из теории [Wang, 2008]. Эксперименты показывают, что она не влияет на качество модели.

Алгоритм 3 Сэмплирование Гиббса LDA-GS.

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

обнулить $\hat{n}_{wt}$, $\hat{n}_{dt}$, $\hat{n}_t$, $\forall d \in D$, $\forall w \in W$, $\forall t \in T$;
повторять
  для всех $d \in D$, $w \in d$, $i = 1, \ldots, n_{dw}$
    если не первый проход коллекции **то**
      $t := t_{dwi}$; уменьшить $\hat{n}_{wt}$, $\hat{n}_{dt}$, $\hat{n}_t$ на 1;
    вычислить ϕ_{wt}, θ_{td} согласно (10);
    сэмплировать t_{dwi} из $p(t \mid d, w) \propto \phi_{wt}\theta_{td}$;
    $t := t_{dwi}$; увеличить $\hat{n}_{wt}$, $\hat{n}_{dt}$, $\hat{n}_t$ на 1;
пока Θ и Φ не стабилизируются.
обновить ϕ_{wt}, θ_{td}, $\forall d \in D$, $\forall w \in W$, $\forall t \in T$;

Можно одновременно уменьшать счетчики для старой темы и увеличивать для новой, как в Алгоритме 2.

4. Единственным существенным различием, влияющим на качество модели, является применение байесовской регуляризации в LDA, которая подробнее рассматривается в следующем параграфе. Различие заключается только в формулах частотных оценок условных вероятностей: в PLSA используются несмещенные оценки максимального правдоподобия (7)–(8), в LDA — байесовские сглаженные оценки (10).

Таким образом, LDA-GS существенно отличается от PLSA-EM только тремя эвристиками: частым обновлением параметров, сэмплированием и регуляризацией. Эти эвристики не связаны друг с другом и могут применяться в любых сочетаниях.

Регуляризация

Тематическая модель LDA [Blei et al., 2003] основана на разложении (2) при дополнительном предположении, что векторы документов $\theta_d \in \mathbb{R}^{|T|}$ и векторы тем $\phi_t \in \mathbb{R}^{|W|}$ порождаются распределениями Дирихле с гиперпараметрами $\alpha = (\alpha_t) \in \mathbb{R}^{|T|}$ и $\beta = (\beta_w) \in \mathbb{R}^{|W|}$ соответственно. Это приводит к сглаженным частотным оценкам [Steyvers, Griffiths, 2004; Wang, 2008]

$$\phi_{wt} = \frac{\hat{n}_{wt} + \beta_w}{\hat{n}_t + \beta_0}, \quad \beta_0 = \sum_{w \in W} \beta_w; \qquad \theta_{td} = \frac{\hat{n}_{dt} + \alpha_t}{n_d + \alpha_0}, \quad \alpha_0 = \sum_{t \in T} \alpha_t. \tag{10}$$

Распределение Дирихле выбрано в качестве априорного по нескольким причинам. Во-первых, оно порождает нормированные случайные векторы, разреженностью которых управляет векторный параметр α (или β) той же размерности. Во-вторых, возникает двухуровневая модель порождения данных, которая хорошо описывает выборки, обладающие кластерной структурой: сначала распределение Дирихле порождает темы ϕ_t — случайные векторы, которые становятся центрами кластеров; затем эти векторы, будучи дискретными распределениями, порождают выборки терминов, эмпирические распределения которых группируются вокруг своих центров; собственно документы формируются путем объединения таких выборок терминов. Эта модель вполне соответствует интуитивному пониманию того, как формируется коллекция документов разной тематики. Наконец, в-третьих, распределение Дирихле является сопряженным к мультиномиальному, что упрощает байесовский вывод, в частности получение формул (10).

В первых работах по LDA [Blei et al., 2003] и сэмплированию Гиббса [Steyvers, Griffiths, 2004] использовались только симметричные распределения Дирихле с равными значениями

всех координат вектора гиперпараметров. Эти значения либо фиксировались, либо настраивались путем перебора по сетке. Позже были предложены эффективные численные методы оптимизации гиперпараметров и экспериментально показано, что их оптимизация улучшает качество модели [Wallach et al., 2009; Wallach, 2008]. При этом координаты вектора гиперпараметров α выгоднее оптимизировать по отдельности, а координаты вектора β лучше брать равными и оптимизировать как один скалярный гиперпараметр.

Известно несколько способов оценивания параметров Θ и Φ в модели LDA, отличающихся, главным образом, формулой сглаживания частотных оценок вероятностей. Сравнение шести наиболее известных способов в [Asuncion et al., 2009] показало, что оптимизация гиперпараметров практически нивелирует различия между ними.

Модель PLSA является частным случаем модели LDA, когда априорные распределения Дирихле вырождаются в равномерные, а для оценивания параметров Φ, Θ применяется метод максимума апостериорной вероятности [Girolami, Kabán, 2003]. Этот случай формально соответствует обнулению всех гиперпараметров в (10): $\alpha_t = 0,\ \beta_w = 0$.

Эксперименты показали, что LDA обеспечивает существенно меньшие значения контрольной перплексии, чем PLSA [Blei et al., 2003]. По аналогии с задачами классификации и регрессии отсюда был сделан стандартный вывод, что модель PLSA имеет слишком много параметров θ_{td}, ϕ_{wt}, на которые не накладывается никаких ограничений, потому возникает переобучение, а в модели LDA эти оценки более устойчивы благодаря байесовской регуляризации, поэтому эффективная сложность модели меньше и переобучение меньше.

Однако более разумной представляется несколько иная интерпретация этих экспериментов. Оптимальные значения гиперпараметров α и β в LDA обычно близки к нулю и могут повлиять только на частотные оценки тем, редких в документе, и терминов, редких в теме. Полезность таких оценок для выявления тематики представляется сомнительной. Контрольная перплексия лучше у LDA только потому, что новым терминам, которых не было в обучающей коллекции, назначаются чуть более адекватные априорные оценки вероятностей $\phi_{wt} = \beta_w/\beta_0$.

Эту гипотезу нетрудно проверить в эксперименте. Если коллекцию разбить на обучающую и контрольную так, чтобы в контрольных документах новых терминов не было, то регуляризация не дает никакого выигрыша и перплексии PLSA и LDA практически совпадают, см. рис. 1. Этот результат согласуется с распространенным мнением, что для больших коллекций нет существенных различий в качестве моделей PLSA и LDA.

Регуляризация создает ряд проблем: необходимо оптимизировать гиперпараметры, инициализировать β_w для новых терминов w и обеспечивать разреженность при том, что обнулять параметры θ_{td} и ϕ_{wt} невозможно. Предположение, что редкие и новые термины бесполезны для тематической модели, приводит к робастным моделям, которым не нужна регуляризация.

Робастная тематическая модель PLSA с компонентами шума и фона

Робастная тематическая модель является обобщением (2). Это вероятностная смесь трех компонент — тематической, шумовой и фоновой [Chemudugunta et al., 2006]:

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

где *шумовая* компонента $\pi_{dw} \equiv p_{\text{ш}}(w\,|\,d)$ — это неизвестное распределение терминов в документе d, *фоновая* компонента $\pi_w \equiv p_{\text{ф}}(w)$ — неизвестное распределение терминов во всей коллекции. Априорные вероятности тематической, шумовой и фоновой компонент равны соответственно $q_{\text{т}} = \frac{1}{1+\gamma+\varepsilon}$, $q_{\text{ш}} = \frac{\gamma}{1+\gamma+\varepsilon}$, $q_{\text{ф}} = \frac{\varepsilon}{1+\gamma+\varepsilon}$, где параметры γ и ε можно как фиксировать, так и оптимизировать.

Чтобы получить приближенное решение М-шага, запишем лагранжиан данной задачи при ограничениях нормировки и неотрицательности π_{dw}, π_w, проигнорировав ограничения неотрицательности θ_{td} и ϕ_{wt}, которые будут выполнены автоматически:

$$\mathcal{L}(D;\Theta,\Phi,\Pi) = \sum_{d\in D}\sum_{w\in d} n_{dw}\ln\frac{Z_{dw}+\gamma\pi_{dw}+\varepsilon\pi_w}{1+\gamma+\varepsilon} + \sum_{d\in D}\sum_{w\in d}\kappa_{dw}\pi_{dw} + \sum_{w\in d}\kappa'_w\pi_w - \\ - \sum_{t\in T}\lambda_t\Big(\sum_{w\in W}\phi_{wt}-1\Big) - \sum_{d\in D}\mu_d\Big(\sum_{t\in T}\theta_{td}-1\Big) - \sum_{d\in D}\nu_d\Big(\sum_{w\in d}\pi_{dw}-1\Big) - \nu'\Big(\sum_{w\in W}\pi_w-1\Big).$$

Двойственные переменные κ_{dw}, соответствующие ограничениям $\pi_{dw}\geqslant 0$, должны быть неотрицательны и удовлетворять условиям дополняющей нежесткости

$$\kappa_{dw}\pi_{dw}=0, \quad d\in D,\ w\in d.$$

Аналогично, для двойственных переменных κ'_w, соответствующих $\pi_w\geqslant 0$:

$$\kappa'_w\pi_w=0, \quad w\in d.$$

По аналогии со стандартным EM-алгоритмом, на Е-шаге для каждой пары (d,w) вычисляются по формуле Байеса условные вероятности тем $H_{dwt}=p(t\,|\,d,w)$:

$$H_{dwt}=\frac{\phi_{wt}\theta_{td}}{Z_{dw}+\gamma\pi_{dw}+\varepsilon\pi_w}, \tag{11}$$

и, кроме того, условные вероятности шума H_{dw} и фона H'_{dw}:

$$H_{dw}=\frac{\gamma\pi_{dw}}{Z_{dw}+\gamma\pi_{dw}+\varepsilon\pi_w}; \qquad H'_{dw}=\frac{\varepsilon\pi_w}{Z_{dw}+\gamma\pi_{dw}+\varepsilon\pi_w}. \tag{12}$$

Продифференцировав лагранжиан по переменным θ_{td} и ϕ_{wt} и приравняв нулю производные, получим прежние формулы для ϕ_{wt} (7) и θ_{td} (8), с единственным отличием, что теперь H_{dwt} вычисляются по новой формуле (11).

Продифференцируем лагранжиан по π_{dw} и приравняем нулю производную:

$$\nu_d=\frac{n_{dw}\gamma}{Z_{dw}+\gamma\pi_{dw}+\varepsilon\pi_w}+\kappa_{dw}. \tag{13}$$

Домножим обе части этого равенства на π_{dw}, просуммируем по всем терминам $w\in W$, применим условие нормировки вероятностей π_{dw} в левой части и условие дополняющей нежесткости в правой части. Получим выражение двойственной переменной ν_d через все основные переменные:

$$\nu_d=\sum_{w\in d} n_{dw}\frac{\gamma\pi_{dw}}{Z_{dw}+\gamma\pi_{dw}+\varepsilon\pi_w}=\sum_{w\in d} n_{dw}H_{dw}. \tag{14}$$

Поскольку H_{dw} есть апостериорная вероятность того, что термин w в документе d является шумом, ν_d интерпретируется как оценка числа шумовых терминов в документе d.

Проделав аналогичные действия для фоновой компоненты, получим

$$\nu'=\frac{n_w\varepsilon}{Z_{dw}+\gamma\pi_{dw}+\varepsilon\pi_w}+\kappa'_w,$$
$$\nu'=\sum_{d\in D}\sum_{w\in d}\frac{n_w\varepsilon\pi_w}{Z_{dw}+\gamma\pi_{dw}+\varepsilon\pi_w}=\sum_{d\in D}\sum_{w\in d} n_{dw}H'_{dw},$$

где ν' интерпретируется как оценка числа фоновых терминов во всей коллекции.

Домножим обе части (13) на π_{dw}, но не будем суммировать по w. Получим формулу М-шага для шумовой компоненты:

$$\pi_{dw} = \frac{1}{\nu_d} n_{dw} \frac{\gamma \pi_{dw}}{Z_{dw} + \gamma \pi_{dw} + \varepsilon \pi_w} = \frac{n_{dw} H_{dw}}{\sum\limits_{w \in d} n_{dw} H_{dw}}.$$

Аналогично получается формула М-шага для фоновой компоненты:

$$\pi_w = \frac{1}{\nu'} n_{dw} \frac{\varepsilon \pi_w}{Z_{dw} + \gamma \pi_{dw} + \varepsilon \pi_w} = \frac{\sum\limits_{d \in D} n_{dw} H'_{dw}}{\sum\limits_{d \in D} \sum\limits_{w \in d} n_{dw} H'_{dw}}.$$

Если начальные приближения π_{dw}, π_w неотрицательны, то они и далее будут неотрицательными на каждом шаге.

Существует альтернативный способ получения формулы для π_{dw}. Перепишем (13) в другом виде:

$$n_{dw}\gamma = (\nu_d - \kappa_{dw})(Z_{dw} + \gamma \pi_{dw} + \varepsilon \pi_w).$$

Согласно условиям дополняющей нежесткости, хотя бы одна из двух неотрицательных переменных κ_{dw}, π_{dw} должна быть равна нулю. Поэтому, если $n_{dw}\gamma < \nu_d(Z_{dw} + \varepsilon \pi_w)$, то $\pi_{dw} = 0$ и $\kappa_{dw} > 0$. Если же имеет место противоположное неравенство, то $\kappa_{dw} = 0$ и π_{dw} находится из уравнения $n_{dw}\gamma = \nu_d(Z_{dw} + \gamma \pi_{dw} + \varepsilon \pi_w)$. Объединяя оба эти случая, получаем выражение для π_{dw} через остальные переменные, не зависящее от H_{dw}:

$$\pi_{dw} = \left(\frac{n_{dw}}{\nu_d} - \frac{Z_{dw} + \varepsilon \pi_w}{\gamma} \right)_+. \tag{15}$$

Таким образом, 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 $\pi_{dw} > 0$.

Робастная модификация итерационного процесса PLSA-GEM показана в Алгоритме 4. Возможны различные варианты этого алгоритма: только с шумовой компонентой ($\varepsilon = 0$), только с фоновой компонентой ($\gamma = 0$), с аддитивным и мультипликативным М-шагом. Для простоты показан вариант с шумом и фоном, обновлением параметров по каждой паре (d, w), без сэмплирования, без регуляризации, с аддитивным М-шагом.

Замечания к Алгоритму 4.

1. Главное отличие от обычного PLSA в том, что теперь n_{dw} вхождений термина w в документ d распределяются не только между темами $t \in T$, но также между шумовой и фоновой компонентами (шаги 10–13), пропорционально вероятностям

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

где Z — нормирующий множитель (см. шаг 9).

2. Вспомогательные переменные n_{dwt}, ν_{dw}, ν'_{dw} нужны для того, чтобы значения счетчиков всегда содержали сумму, накопленную при последнем проходе всей коллекции, в соответствии с (7)–(8).

3. Счетчики ν_d, ν, ν'_w, ν' имеют прозрачную содержательную интерпретацию: ν_d, ν — это оценки числа шумовых токенов в документе d и во всей коллекции; ν'_w, ν' — оценки числа фоновых токенов термина w и всех фоновых токенов в коллекции.

4. *Регуляризация* вводится в Алгоритм 4 заменой частотных оценок (7)–(8) параметров ϕ_{wt}, θ_{td} на шагах 5, 6, 15 сглаженными оценками (10).

Алгоритм 4 Робастный PLSA-GEM.

Вход: коллекция D, число тем $|T|$, начальные приближения Θ, Φ, параметры γ, ε;
Выход: распределения Φ, Θ, Π;

1: инициализировать $\forall d \in D,\ \forall w \in W,\ \forall t \in T$:
$\hat{n}_{wt}, \hat{n}_{dt}, \hat{n}_t, \hat{n}_d, n_{dwt}, \nu_{dw}, \nu_d, \nu, \nu'_{dw}, \nu'_w, \nu' := 0$;
$\pi_{dw} := n_{dw}/n_d;\ \ \pi_w := n_w/n$;
2: **повторять**
3: **для всех** $d \in D,\ w \in d$
4: **если** не первый проход коллекции **то**
5: $\phi_{wt} := \hat{n}_{wt}/\hat{n}_t;\ \ \forall t \in T$;
6: $\theta_{td} := \hat{n}_{dt}/\hat{n}_d;\ \ \forall t \in T$;
7: $\pi_w := \nu'_w/\nu'$;
8: $\pi_{dw} := (n_{dw}/\nu_d - Z_{dw}/\gamma - \varepsilon\pi_w/\gamma)_+$;
9: $Z := Z_{dw} + \gamma\pi_{dw} + \varepsilon\pi_w$;
10: **для всех** $t \in T$: $n_{dwt} > 0$ или $\phi_{wt}\theta_{td} > 0$
11: $\delta_т := n_{dw}\phi_{wt}\theta_{td}/Z$; увеличить $\hat{n}_{wt}, \hat{n}_{dt}, \hat{n}_t, \hat{n}_d$ на $(\delta_т - n_{dwt})$; $n_{dwt} := \delta_т$;
12: $\delta_ш := n_{dw}\gamma\pi_{dw}/Z$; увеличить ν_d, ν на $(\delta_ш - \nu_{dw})$; $\nu_{dw} := \delta_ш$;
13: $\delta_ф := n_{dw}\varepsilon\pi_w/Z$; увеличить ν'_w, ν' на $(\delta_ф - \nu'_{dw})$; $\nu'_{dw} := \delta_ф$;
14: **пока** Φ, Θ, Π не стабилизируются.
15: обновить $\phi_{wt}, \theta_{td}, \pi_w, \pi_{dw}$ для всех d, w, t;

5. *Сэмплирование* вводится заменой распределения $\tilde{H}_{dw}$ его эмпирической оценкой, аналогичной (9), при вычислении переменных $\delta_т$, $\delta_ш$, $\delta_ф$ (шаги 11, 12, 13).

6. В экспериментах аддитивный шаг всегда давал лучшее качество и сходимость.

7. Попытка оптимизировать параметры γ и ε приводит к тому, что тематическая модель вырождается в шумовую компоненту ($\gamma \to \infty,\ \ \varepsilon \to 0$), т. е. для каждого документа строится отдельная униграммная модель [Blei et al., 2003]. Поэтому параметры γ и ε необходимо фиксировать.

Разреженность

Гипотеза разреженности предполагает, что в дискретных распределениях $p(w|t) = \phi_{wt}$, $p(t|d) = \theta_{td}$, $p(t|d,w) = H_{dwt}$ подавляющее большинство вероятностей равны нулю или очень близки к нулю. Алгоритмы, в которых нулевые значения не хранятся, намного эффективнее по памяти и по скорости. Поэтому для больших коллекций разреженность обязательна. Недостаток обнуления в том, что при $p(w|d) = 0$ и $n_{dw} > 0$ трудно адекватно оценить качество модели — перплексия в этом случае уходит в бесконечность.

Модель PLSA исходно не является разреженной. Согласно шагам 5–8 Алгоритма 1, если $\theta_{td} = 0$ (тема t не представлена в документе d) или если $\phi_{wt} = 0$ (термин w не относится к теме t), то нулевое значение будет сохраняться на протяжении всех итераций. И, наоборот, если значения θ_{td}, ϕ_{wt} положительны в начальном приближении, то они так и останутся положительными. Таким образом, PLSA не оптимизирует структуру разреженности распределений и требует задавать ее через начальное приближение. Отдельные значения θ_{td} и ϕ_{wt} могут в ходе итераций стремиться к нулю, но, как правило, их доля недостаточна для получения выигрыша в производительности. Принудительное разреживание путем обнуления малых значений θ_{td} и ϕ_{wt} с последующей перенормировкой может приводить к ухудшению качества модели, особенно на первых итерациях.

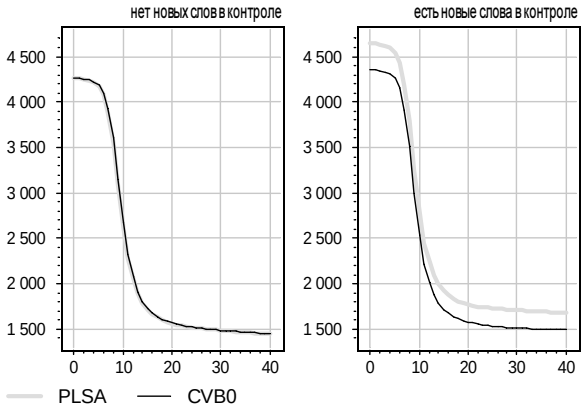

Рис. 1. Регуляризация дает преимущество, только когда в контроле есть новые термины (метод CVB0 — это PLSA-GEM с регуляризацией, но без сэмплирования)

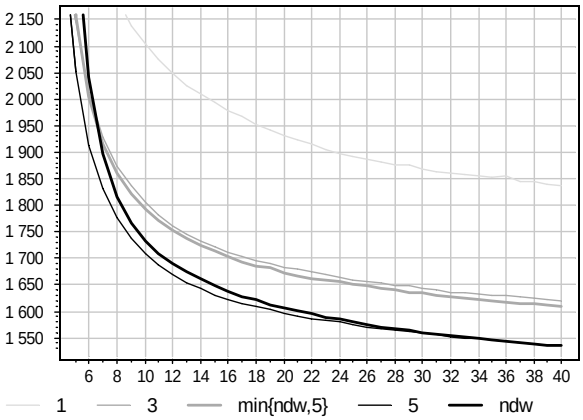

Рис. 2. При экономном сэмплировании пяти тем для каждой пары (d, w) перплексия не хуже, чем при сэмплировании n_{dw} тем. Но одной или трех тем недостаточно

Модель LDA также не является разреженной — сглаживание частотных оценок вероятностей приводит к тому, что матрицы Φ и Θ не содержат нулевых значений. Эта проблема имеет несколько возможных решений, например в [Eisenstein et al., 2011] предлагается хранить не сами значения θ_{td} и ϕ_{wt}, а только их разности с фоновыми распределениями.

Робастные модели допускают разреженность тематической компоненты модели и одновременно исключают ситуацию бесконечной перплексии, так как нулевое значение Z_{dw} компенсируется ненулевым значением шумовой компоненты $p_{\text{ш}}(w \mid d)$. Чем больше γ, тем более разреженной может быть тематическая компонента модели.

Эвристика принудительного разреживания. В эксперименте с робастным PLSA на каждой итерации принудительно обнулялись 5 % наименьших значений θ_{td} и ϕ_{wt}. При этом разреженность матриц Θ и Φ достигала порядка 90 % без существенной потери качества модели (рис. 4).

Эксперименты на реальных данных

Обучение модели производилось по коллекции из $|D| = 2000$ авторефератов диссертаций на русском языке суммарной длины $n \approx 8.7 \cdot 10^6$, объем словаря $|W| \approx 3 \cdot 10^4$.

Предварительно производилась лемматизация и отбрасывались стоп-слова.

Качество модели оценивалось функционалом *перплексии контрольных данных* (4) по контрольной коллекции D' из 200 авторефератов, не включенных в обучающую коллекцию. Каждый контрольный документ d случайным образом делится на две половины, d' и d''. Параметры θ_{td} и ν_d оцениваются по d'. Параметры ϕ_{wt} и π_w оцениваются по обучающей выборке D. Параметры π_{dw} оцениваются для каждой пары (d, w) согласно (15). Перплексия вычисляется по вторым половинам d'' контрольных документов.

На рис. 1–4 показаны зависимости перплексии от числа итераций (одна итерация — это один проход по коллекции). Число итераций 40; число тем $|T| = 100$; параметры регуляризации $\alpha_t = 0.5$, $\beta_w = 0.01$; параметры робастности $\gamma = 0.3$, $\varepsilon = 0.1$.

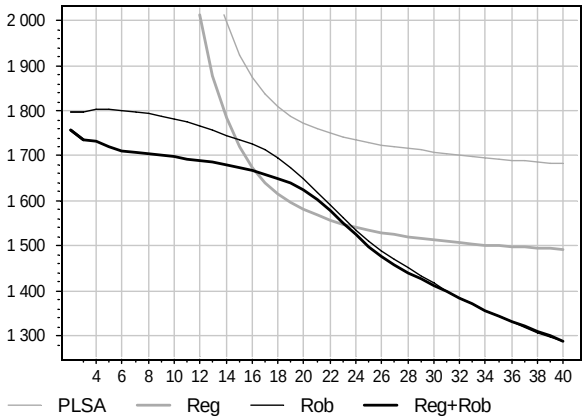

Рис. 3. Робастность сильнее уменьшает перплексию PLSA, чем регуляризация. Регуляризация не улучает робастную модель.

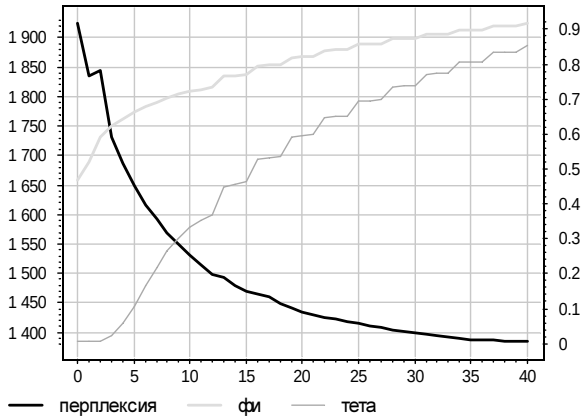

Рис. 4. В процессе разреживания доля нулевых ϕ_{wt} и θ_{td} (отложена по правой оси) увеличивается при монотонном уменьшении перплексии.

Список литературы

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

Asuncion A., Welling M., Smyth P., Teh Y. W. On smoothing and inference for topic models // Int'l conf. on Uncertainty in Artificial Intelligence. — 2009.

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

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.

Daud A., Li J., Zhou L., Muhammad F. Knowledge discovery through directed probabilistic topic models: a survey // Frontiers of Computer Science in China. — 2010. — Vol. 4, no. 2. — Pp. 280–301.

Dempster A. P., Laird N. M., Rubin D. B. Maximum likelihood from incomplete data via the EM algorithm // Journal of the Royal Statistical Society, Series B. — 1977. — no. 34. — Pp. 1–38.

Eisenstein J., Ahmed A., Xing E. P. Sparse additive generative models of text // International Conference on Machine Learning, ICML'11. — 2011. — Pp. 1041–1048.

Girolami M., Kabán A. On an Equivalence between PLSI and LDA // Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval. — ACM, 2003. — Pp. 433–434.

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.

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

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

Wallach H., Mimno D., McCallum A. Rethinking LDA: Why priors matter // Advances in Neural Information Processing Systems 22 — 2009. — Pp. 1973–1981.

Wang Y. Distributed Gibbs sampling of latent Dirichlet allocation: The gritty details. — 2008. `dbgroup.cs.tsinghua.edu.cn/wangyi/lda/lda.pdf`.