
Аддитивная регуляризация тематических моделей

Воронцов К. В. · Потапенко А. А.

20 января 2014

Вероятностное тематическое моделирование коллекций текстовых документов развивается в настоящее время, главным образом, в рамках байесовского подхода и графических моделей. В данной работе предлагается альтернативный подход, свободный от избыточных вероятностных предположений. Аддитивная регуляризация тематических моделей (ARTM) основана на максимизации взвешенной суммы логарифма правдоподобия и дополнительных критериев — регуляризаторов. Это упрощает комбинирование тематических моделей и построение сколь угодно сложных многоцелевых моделей. Многие известные модели рассматриваются как регуляризаторы в терминах ARTM. В экспериментах на реальных данных исследуются комбинации регуляризаторов сглаживания, разреживания и декорреляции. Вместе они позволяют существенно улучшить критерии разреженности, когерентности, чистоты и контрастности тем, практически без ухудшения точности тематической модели.

Ключевые слова: вероятностное тематическое моделирование · регуляризация некорректно поставленных обратных задач · вероятностный латентный семантический анализ · латентное размещение Дирихле

1. Введение

Тематическое моделирование — одно из активно развивающихся направлений статистического анализа текстов [1]. *Вероятностная тематическая модель* выявляет тематику коллекции текстовых документов, описывая каждую тему дискретным распределением на множестве терминов, каждый документ — дискретным распределением на множестве тем. Тематические модели используются для ин-

Воронцов Константин Вячеславович
Московский Физико-Технический Институт, Высшая Школа Экономики, Яндекс,
Вычислительный Центр им. А. А. Дородницына РАН, E-mail: voron@yandex-team.ru

Потапенко Анна Александровна
Московский Государственный Университет им. М. В. Ломоносова,
Вычислительный Центр им. А. А. Дородницына РАН, E-mail: anya_potapenko@mail.ru

формационного поиска, классификации, категоризации, аннотирования, сегментации текстов.

Латентное размещение Дирихле LDA [2] является доминирующим подходом в вероятностном тематическом моделировании. На базе LDA разработаны сотни специализированных моделей. Двухуровневая порождающая модель LDA основана на предположении, что распределения терминов в темах и тем в документах — это нормированные векторы, порождаемые распределениями Дирихле. Распределение Дирихле является сопряжённым к дискретному распределению, что существенно упрощает байесовский вывод. В то же время, применение байесовского подхода в тематическом моделировании не лишено недостатков. Распределение Дирихле не имеет убедительных лингвистических обоснований и формально не допускает разреженных решений. Байесовский вывод затрудняет построение многоцелевых тематических моделей, удовлетворяющих одновременно большому числу дополнительных требований.

В данной работе мы предлагаем альтернативу байесовскому подходу — *аддитивную регуляризацию тематических моделей*, ARTM. Это приложение классической теории регуляризации некорректно поставленных задач [3] к тематическому моделированию. Обычно построение тематической модели сводится к задаче стохастического матричного разложения. В общем случае она имеет бесконечно много решений, то есть является некорректно поставленной. Для её регуляризации к логарифму правдоподобия добавляются штрафные слагаемые, формализующие дополнительные требования к модели.

ARTM имеет несколько принципиальных отличий от байесовского подхода.

Во-первых, не ставится задача построения чисто вероятностной модели порождения текста. Многие лингвистические ограничения легче формализуются с помощью оптимизационных критериев, а не через априорные распределения. При этом регуляризаторы не обязаны иметь вероятностную интерпретацию. Распределение Дирихле утрачивает роль «главного регуляризатора» и уступает место разнообразным проблемно-ориентированным регуляризаторам. Структура моделей становится настолько очевидной, что отпадает необходимость её пояснения на языке графических моделей.

Во-вторых, вместо байесовского вывода используется более простой подход — регуляризованный ЕМ-алгоритм. ARTM не требует интегрирования по пространству параметров модели и сводится к дифференцированию регуляризаторов по параметрам. Построение многоцелевых тематических моделей [4] существенно упрощается благодаря аддитивности регуляризаторов. Добавление регуляризатора требует небольшой модификации М-шага в готовом ЕМ-подобном алгоритме.

ARTM отличается также и от ранее предлагавшихся методов регуляризации [5, 6, 7, 8]. В каждом из них использовался какой-либо конкретный регуляризатор: KL-дивергенция, распределение Дирихле, L_1 - или L_2 -норма. ARTM — это не инкрементное улучшение одной тематической модели, а общий подход к тематическому моделированию как к задаче многокритериальной оптимизации.

Цель данной работы — предложить общие методы регуляризации сложных многоцелевых тематических моделей и описать начальный набор регуляризаторов, полезных для решения многих прикладных задач.

В разделе 2 мы рассматриваем модель PLSA, которая исторически предшествовала LDA. Она не имеет регуляризаторов, поэтому является наиболее удобной платформой для ARTM. Мы приводим ЕМ-алгоритм и его элементарное обосно-

вание, описываем эксперимент на модельных данных, демонстрирующий неединственность и неустойчивость PLSA и LDA.

В разделе 3 мы вводим аддитивную регуляризацию тематических моделей и обосновываем формулу регуляризованного М-шага.

В разделе 4 мы показываем, что многие известные тематические модели легко выводятся с помощью ARTM. В частности, модель LDA выводится из регуляризатора, минимизирующего дивергенцию Кульбака–Лейблера с фиксированным распределением. Рассматриваются регуляризаторы для сглаживания, разреживания, частичного обучения, декорреляции, выявления коррелированных тем, улучшения когерентности, классификации документов и балансировки классов.

В разделе 5 мы иллюстрируем применение ARTM на примере комбинирования разреживания, сглаживания и декорреляции тем с целью выделения стоп-слов и улучшения интерпретируемости тематической модели. Эксперименты показывают, что комбинирование регуляризаторов приводит к улучшению тематической модели по совокупности критериев.

В разделе 6 приводятся интерпретации результатов и основные выводы.

2. Тематические модели PLSA и LDA

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, \dots, 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) , $i = 1, \dots, n$ 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.

Гипотеза независимости или «мешка слов» означает, что тематику документа можно узнать даже после произвольной перестановки терминов, хотя для человека такой текст теряет смысл. Это предположение позволяет перейти к компактному представлению документа как подмножества $d \subset W$, в котором каждому элементу $w \in d$ поставлено в соответствие число n_{dw} вхождений термина w в документ d .

Гипотезой *условной независимости* называется предположение, что появление слов по теме t не зависит от документа: $p(w | t) = p(w | d, t)$. Согласно формуле полной вероятности и гипотезе условной независимости

$$p(w | d) = \sum_{t \in T} p(w | t) p(t | d). \quad (1)$$

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

Обычно число тем $|T|$ много меньше $|D|$ и $|W|$, и задача сводится к поиску приближённого представления заданной матрицы частот

$$F = (\hat{p}_{wd})_{W \times D}, \quad \hat{p}_{wd} = \hat{p}(w | d) = \frac{n_{dw}}{n_d},$$

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

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

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

В *вероятностном латентном семантическом анализе* PLSA [9] для построения модели (1) максимизируется логарифм правдоподобия при ограничениях нормировки и неотрицательности:

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

$$\sum_{w \in W} \phi_{wt} = 1, \quad \phi_{wt} \geq 0; \quad \sum_{t \in T} \theta_{td} = 1, \quad \theta_{td} \geq 0. \quad (3)$$

Теорема 1. Стационарная точка оптимизационной задачи (2), (3) удовлетворяет системе уравнений со вспомогательными переменными p_{tdw} , n_{wt} , n_{td} , n_t , n_d

$$p_{tdw} = \frac{\phi_{wt} \theta_{td}}{\sum_{s \in T} \phi_{ws} \theta_{sd}}; \quad (4)$$

$$\phi_{wt} = \frac{n_{wt}}{n_t}, \quad n_{wt} = \sum_{d \in D} n_{dw} p_{tdw}, \quad n_t = \sum_{w \in W} n_{wt}; \quad (5)$$

$$\theta_{td} = \frac{n_{td}}{n_d}, \quad n_{td} = \sum_{w \in d} n_{dw} p_{tdw}, \quad n_d = \sum_{t \in T} n_{td}. \quad (6)$$

Это утверждение следует из условий Куна–Таккера. Далее мы докажем более общее утверждение. Система уравнений (4)–(6) может быть решена различными численными методами. В частности, метод простых итераций приводит к ЕМ-алгоритму, который чаще всего используется на практике.

На каждой итерации ЕМ-алгоритма выполняются два шага.

Е-шаг (4) можно интерпретировать как применение формулы Байеса для вычисления условных распределений $p(t|d, w)$:

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

Зная эти условные вероятности, можно оценить число вхождений термина w в документ d , связанных с темой t : $n_{tdw} = n_{dw} p_{tdw}$.

На М-шаге (5), (6) суммирование n_{tdw} по d, w, t даёт частотные оценки максимального правдоподобия для искомых условных вероятностей ϕ_{wt} , θ_{td} . Иногда их записывают кратко через знак пропорциональности: $\phi_{wt} \propto n_{wt}$, $\theta_{td} \propto n_{td}$.

В Алгоритме 2.1 ЕМ-итерации организованы так, чтобы Е-шаг вычислялся внутри М-шага. Это позволяет избежать хранения трёхмерного массива p_{tdw} . Каждая ЕМ-итерация — это один проход по коллекции документов.

В модели *латентного размещения Дирихле* LDA [2] вводятся ограничения на параметры Φ, Θ , чтобы избежать переобучения. Предполагается, что столбцы матриц Φ, Θ порождаются распределениями Дирихле с гиперпараметрами, соответственно, $\beta = (\beta_w)_{w \in W}$ и $\alpha = (\alpha_t)_{t \in T}$. Это приводит к различным

Алгоритм 2.1. Рациональный EM-алгоритм для модели PLSA.

Вход: коллекция документов D , число тем $|T|$;
Выход: Φ , Θ ;

- 1 инициализировать вектор-столбцы ϕ_t , θ_d случайным образом;
- 2 **повторять**
 - 3 обнулить n_{wt} , n_{td} , n_t , n_d для всех $d \in D$, $w \in W$, $t \in T$;
 - 4 **для всех** $d \in D$, $w \in d$
 - 5 | $Z := \sum_{t \in T} \phi_{wt} \theta_{td}$;
 - 6 | **для всех** $t \in T$: $\phi_{wt} \theta_{td} > 0$
 - 7 | | увеличить n_{wt} , n_{td} , n_t , n_d на $\delta = n_{dw} \phi_{wt} \theta_{td} / Z$;
 - 8 $\phi_{wt} := n_{wt} / n_t$ for all $w \in W$, $t \in T$;
 - 9 $\theta_{td} := n_{td} / n_d$ for all $d \in D$, $t \in T$;
- 10 **пока** Θ и Φ не сойдутся;

ЕМ-подобным алгоритмам с модифицированной формулой М-шага [10], из которых чаще всего используется наиболее простая:

$$\phi_{wt} \propto n_{wt} + \beta_w, \quad \theta_{td} \propto n_{td} + \alpha_t. \quad (7)$$

В алгоритме сэмплирования Гиббса [11] для каждого вхождения термина w в документ d выбирается случайная тема из распределения $p(t | d, w)$, счётчики числа слов n_{wt} , n_{td} , n_t , n_d увеличиваются на единицу, после чего новые значения ϕ_{wt} , θ_{td} вычисляются «на лету» по формулам (7).

На больших коллекциях точность моделей PLSA и LDA отличается незначительно [12, 13, 14, 15]. Оптимальные значения параметров β_w , α_t обычно близки к нулю [16], поэтому добавки β_w , α_t в (7) существенны только при малых n_{wt} , n_{td} . Применение робастных вариантов PLSA и LDA, игнорирующих редкие термины в темах и редкие темы в документах, нивелирует различия между PLSA и LDA [15]. Таким образом, LDA снижает переобучение только на редких терминах и темах, незначимых для тематической модели. На больших данных переобучение не является проблемой для обеих моделей.

Действительно серьёзной проблемой, редко обсуждаемой в литературе, является неединственность 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 ЕМ-подобных алгоритмах никак не контролируется и зависит от случайного начального приближения.

Чтобы исследовать способность PLSA и LDA восстанавливать истинные матрицы Φ и Θ , был проведён эксперимент на модельной коллекции при $|W| = 1000$, $|D| = 500$, $|T| = 30$ и случайных длинах документов $n_d \in [100, 600]$. Столбцы матриц Φ , Θ генерировались из симметричных распределений Дирихле с параметрами α , β соответственно. Отклонение восстановленных распределений $\hat{p}(i | j)$ от модельных $p(i | j)$ измерялось средним расстоянием Хеллингера как для самих матриц Φ , Θ , так и для их произведения:

$$D_\Phi = H(\hat{\Phi}, \Phi); \quad D_\Theta = H(\hat{\Theta}, \Theta); \quad D_{\Phi\Theta} = H(\hat{\Phi}\hat{\Theta}, \Phi\Theta);$$

$$H(\hat{p}, p) = \frac{1}{m} \sum_{j=1}^m \sqrt{\frac{1}{2} \sum_{i=1}^n \left(\sqrt{\hat{p}(i | j)} - \sqrt{p(i | j)} \right)^2}.$$

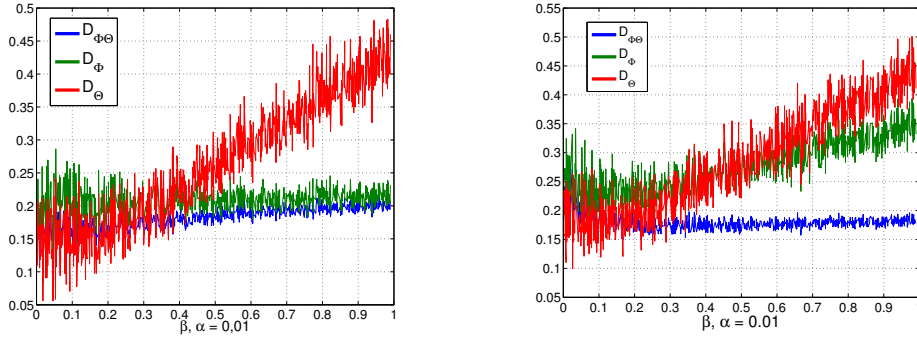


Рис. 1. Зависимость ошибки восстановления матриц Φ , Θ и $\Phi\Theta$ от гиперпараметра β при фиксированном $\alpha = 0.01$, для алгоритмов LDA-GS (Gibbs Sampling) и PLSA-EM.

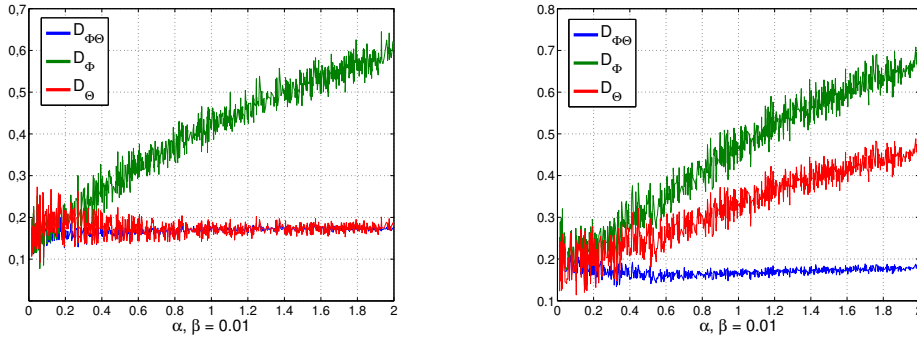


Рис. 2. Зависимость ошибки восстановления матриц Φ , Θ и $\Phi\Theta$ от гиперпараметра α при фиксированном $\beta = 0.01$, для алгоритмов LDA-GS (Gibbs Sampling) и PLSA-EM.

Оказалось, что матрицы Φ , Θ восстанавливаются гораздо хуже, чем их произведение, рис. 1, 2. Ошибка восстановления зависит от разреженности исходных матриц Φ , Θ и не устраняется даже в LDA, если брать те же гиперпараметры α , β , при которых генерировались модельные данные.

Таким образом, распределение Дирихле — слишком слабый регуляризатор. Для обеспечения устойчивости решения необходимы проблемно-ориентированные регуляризаторы, формализующие более сильные предположения о матрицах Φ , Θ . Далее мы берём за основу модель PLSA, свободную от регуляризаторов, отказываясь от «регуляризатора по умолчанию», навязываемого моделью LDA.

3. Аддитивная регуляризация ARTM

Допустим, что наряду с правдоподобием (2) требуется максимизировать ещё r критериев $R_i(\Phi, \Theta)$, $i = 1, \dots, r$, называемых *регуляризаторами* [3]. Для многокритериальной оптимизации будем максимизировать линейную комбинацию критериев $L(\Phi, \Theta)$ и $R_i(\Phi, \Theta)$ с неотрицательными *коэффициентами регуляризации* τ_i :

$$R(\Phi, \Theta) = \sum_{i=1}^r \tau_i R_i(\Phi, \Theta), \quad L(\Phi, \Theta) + R(\Phi, \Theta) \rightarrow \max_{\Phi, \Theta}. \quad (8)$$

Тема t называется *регулярной*, если $n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} > 0$ хотя бы для одного термина $w \in W$, иначе будем говорить, что тема t *перерегуляризована*.

Документ d называется *регулярным*, если $n_{td} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} > 0$ хотя бы для одной темы $t \in T$, иначе будем говорить, что документ d *перерегуляризован*.

Теорема 2. Если функция $R(\Phi, \Theta)$ непрерывно дифференцируема и (Φ, Θ) — точка локального экстремума задачи (8), (3), то для всех регулярных тем t и регулярных документов d справедлива система уравнений:

$$p_{tdw} = \frac{\phi_{wt}\theta_{td}}{\sum_{s \in T} \phi_{ws}\theta_{sd}}; \quad (9)$$

$$\phi_{wt} \propto \left(n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right)_+; \quad n_{wt} = \sum_{d \in D} n_{dw} p_{tdw}; \quad (10)$$

$$\theta_{td} \propto \left(n_{td} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right)_+; \quad n_{td} = \sum_{w \in d} n_{dw} p_{tdw}; \quad (11)$$

где $(z)_+ = \max\{z, 0\}$.

Замечание 1. Если тема t не регулярна, то из (10) следует $\phi_t = 0$. В этом случае мы исключаем тему t из модели. Перерегуляризация тем — это полезный механизм, позволяющий удалять нерелевантные темы и оптимизировать число тем.

Замечание 2. Если документ d не регулярный, то из (11) следует $\theta_d = 0$. В этом случае мы исключаем документ d из модели. Например, документ может быть слишком коротким или не иметь отношения к тематике коллекции.

Замечание 3. Теорема 1 является частным случаем Теоремы 2 при $R(\Phi, \Theta) = 0$.

Доказательство. Запишем необходимые условия локального экстремума (Φ, Θ) задачи (8), (3) по тереме Куна–Таккера:

$$\sum_d n_{dw} \frac{\theta_{td}}{p(w|d)} + \frac{\partial R}{\partial \phi_{wt}} = \lambda_t - \lambda_{wt}; \quad \lambda_{wt} \geq 0; \quad \lambda_{wt} \phi_{wt} = 0,$$

где λ_t and λ_{wt} — множители Лагранжа соответственно для ограничений нормировки и неотрицательности. Умножим обе части первого равенства на ϕ_{wt} , в левой части выделим вспомогательную переменную p_{tdw} из (9) и просуммируем по d :

$$\phi_{wt} \lambda_t = \sum_d n_{dw} \frac{\phi_{wt} \theta_{td}}{p(w|d)} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} = n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}}.$$

Предположение, что $\lambda_t \leq 0$ противоречит условию регулярности темы t . Следовательно, $\lambda_t > 0$, $\phi_{wt} \geq 0$. Левая часть уравнения неотрицательная, значит, правая часть также неотрицательна, и

$$\phi_{wt} \lambda_t = \left(n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right)_+. \quad (12)$$

Просуммируем обе части уравнения по всем терминам $w \in W$:

$$\lambda_t = \sum_{w \in W} \left(n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right)_+. \quad (13)$$

Наконец, получим (10), выражая ϕ_{wt} из (12) и (13).

Уравнения для θ_{td} выводятся аналогично. Теорема доказана.

Система уравнений (9)–(11) определяет регуляризованный ЕМ-алгоритм. В нём сохраняется Е-шаг (4), а формулы М-шага заменяются регуляризованными уравнениями (10)–(11). Таким образом, ЕМ-алгоритм для обучения регуляризованной модели может быть реализован путём незначительной модификации любого имеющегося ЕМ-подобного алгоритма. В частности, в Алгоритме 2.1 достаточно заменить шаги 8 and 9 в соответствии с уравнениями (10)–(11).

4. Примеры регуляризаторов: обзор тематических моделей

В данном разделе пересматриваются тематические модели, ранее разработанные в рамках байесовского подхода. Для каждой из них удаётся найти соответствующий регуляризатор, который по Теореме 2 приводит к тому же самому или очень похожему алгоритму обучения модели. По сравнению с байесовским подходом, ARTM радикально упрощает вывод алгоритма и позволяет комбинировать регуляризаторы в произвольных сочетаниях.

Мы будем использовать дивергенцию Кульбака–Лейблера (относительную энтропию) как меру различия двух дискретных распределений $(p_i)_{i=1}^n$ и $(q_i)_{i=1}^n$:

$$\text{KL}(p||q) \equiv \text{KL}_i(p_i||q_i) = \sum_{i=1}^n p_i \ln \frac{p_i}{q_i}.$$

Минимизация KL-дивергенции эквивалентна максимизации правдоподобия модели распределения q по эмпирическому распределению p .

Сглаживающий регуляризатор и модель LDA. Потребуем, чтобы распределения ϕ_t и θ_d были близки по дивергенции Кульбака–Лейблера к заданным распределениям $\beta = (\beta_w)_{w \in W}$ и $\alpha = (\alpha_t)_{t \in T}$ соответственно:

$$\sum_{t \in T} \text{KL}_w(\beta_w || \phi_{wt}) \rightarrow \min_{\Phi}, \quad \sum_{d \in D} \text{KL}_t(\alpha_t || \theta_{td}) \rightarrow \min_{\Theta}.$$

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

$$R(\Phi, \Theta) = \beta_0 \sum_{t \in T} \sum_{w \in W} \beta_w \ln \phi_{wt} + \alpha_0 \sum_{d \in D} \sum_{t \in T} \alpha_t \ln \theta_{td} \rightarrow \max.$$

Применение общих формул (10) и (11) даёт то же выражение (7) для М-шага, что и модель LDA:

$$\phi_{wt} \propto n_{wt} + \beta_0 \beta_w, \quad \theta_{td} \propto n_{td} + \alpha_0 \alpha_t,$$

если в качестве векторов гиперпараметров взять дискретные распределения β и α , умноженные на коэффициенты регуляризации: $(\beta_0 \beta_t)_{t \in T}$, $(\alpha_0 \alpha_w)_{w \in W}$.

Интерпретация регуляризатора через KL-дивергенцию представляется не менее естественной, чем через априорное распределение Дирихле.

Разреживающий регуляризатор. Предположим, что каждый документ и каждый термин связан с небольшим числом тем. Тогда среди вероятностей ϕ_{wt} и θ_{td} должно быть много нулевых. При построении тематических моделей больших коллекций с большим числом тем сильная разреженность матриц Φ, Θ помогает сократить затраты памяти и времени.

Чем сильнее разрежено распределение, тем меньше его энтропия. Максимальной энтропией обладает равномерное распределение. Поэтому будем максимизировать KL-дивергенцию между модельными распределениями ϕ_t, θ_d и заданными распределениями $\beta = (\beta_w)_{w \in W}, \alpha = (\alpha_t)_{t \in T}$, например, равномерными:

$$R(\Phi, \Theta) = -\beta_0 \sum_{t \in T} \sum_{w \in W} \beta_w \ln \phi_{wt} - \alpha_0 \sum_{d \in D} \sum_{t \in T} \alpha_t \ln \theta_{td} \rightarrow \max.$$

Формулы М-шага, согласно (10) и (11), отличаются от сглаживающего регуляризатора знаком параметра и приводят к разреживанию распределений:

$$\phi_{wt} \propto (n_{wt} - \beta_0 \beta_w)_+, \quad \theta_{td} \propto (n_{td} - \alpha_0 \alpha_t)_+.$$

Идея энтропийной регуляризации была предложена в динамической тематической модели PLSA для разреживания распределений тем во времени при обработке видеопотоков [17]. В данной задаче документами являются видеозаписи, терминами — признаки на изображениях, темами — появление определённого объекта в течение определённого времени, например, проезд автомобиля. Однако возможность применения этой же техники для разреживания распределений ϕ_t и θ_d осталась незамеченной.

Многие исследования, направленные на разреживание модели LDA, приводят к чрезмерно сложным конструкциям [18, 19, 20, 8, 21], поскольку существует внутреннее противоречие между требованием разреженности и свойством распределения Дирихле не допускать нулевых вероятностей. Наш подход к разреживанию намного проще и естественнее. Заметим также, что сглаживание и разреживание описываются одинаково, если не вводить ограничений на знаки параметров β_w, α_t .

Сглаживающий регуляризатор для частичного обучения. Для улучшения интерпретируемости тематической модели могут задаваться обучающие данные. Пусть для документов $d \in D_0$ известно, что они относятся к темам $T_d \subset T$, для тем $t \in T_0$ известно, что к ним относятся термины $W_t \subset W$. Введём регуляризатор, минимизирующий сумму KL-дивергенций между ϕ_{wt} и равномерными распределениями на подмножествах терминов $\beta_{wt} = \frac{1}{|W_t|} [w \in W_t]$, а также между θ_{td} и равномерными распределениями на подмножествах тем $\alpha_{td} = \frac{1}{|T_d|} [t \in T_d]$:

$$R(\Phi, \Theta) = \beta_0 \sum_{t \in T_0} \sum_{w \in W} \beta_{wt} \ln \phi_{wt} + \alpha_0 \sum_{d \in D_0} \sum_{t \in T} \alpha_{td} \ln \theta_{td} \rightarrow \max.$$

Формулы М-шага, согласно (10) и (11), принимают вид

$$\begin{aligned} \phi_{wt} &\propto n_{wt} + \beta_0 \beta_{wt} [t \in T_0]; \\ \theta_{td} &\propto n_{td} + \alpha_0 \alpha_{td} [d \in D_0]. \end{aligned}$$

Это тоже вариант сглаживания, и ещё одно обобщение LDA, но теперь векторы β, α различны для распределений ϕ_t, θ_d и зависят от обучающих данных.

Декоррелирующий регуляризатор для тем. Считается, что повышение различности тем улучшает интерпретируемость модели [22]. Регуляризатор, минимизирующий ковариации между вектор-столбцами ϕ_t, ϕ_s ,

$$R(\Phi) = -\gamma \sum_{t \in T} \sum_{s \in T \setminus t} \sum_{w \in W} \phi_{wt} \phi_{ws} \rightarrow \max,$$

приводит к формуле М-шага

$$\phi_{wt} \propto \left(n_{wt} - \gamma \phi_{wt} \sum_{s \in T \setminus t} \phi_{ws} \right)_+.$$

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. Кроме того, он обладает дополнительным полезным свойством группировать стоп-слова в отдельные темы [22].

Ковариационный регуляризатор для документов. Иногда имеется дополнительная информация о связях между документами схожей тематики. В частности, они могут относиться к одной рубрике, часто совместно упоминаться или ссылаться друг на друга. Формализуем это предположение с помощью регуляризатора:

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

где n_{dc} — вес связи между документами, например, число ссылок на c из d . В [23] предложена похожая модель LDA-JS, в которой вместо максимизации ковариации минимизируется дивергенция Йенсена-Шеннона между θ_d и θ_c . Формула М-шага для θ_{td} , согласно (11), принимает вид

$$\theta_{td} \propto n_{td} + \tau \theta_{td} \sum_{c \in D} n_{dc} \theta_{tc}.$$

Это ещё одна разновидность сглаживания. Вероятности θ_{td} в ходе итераций приближаются к вероятностям θ_{tc} документов, связанных с d .

Коррелированная тематическая модель (Correlated Topic Model, СТМ) предназначена для выявления корреляций между темами [24]. Например, статья по геологии более вероятно связана с археологией, чем с генетикой. Коррелированная модель оценивает корреляции между координатами векторов документов θ_d . Это позволяет точнее моделировать тематику новых документов, повышая вероятности тем, которые часто совместно встречаются в документах коллекции.

Для описания корреляций удобно использовать многомерное нормальное распределение. Однако вектор-столбцы θ_d описывают разреженные дискретные распределения и не похожи на нормальные векторы. Поэтому в модель вводится многомерное лог-нормальное распределение (logistic normal):

$$\theta_{td} = \frac{\exp(\eta_{td})}{\sum_{s \in T} \exp(\eta_{sd})}; \quad p(\eta_d | \mu, \Sigma) = \frac{\exp(-\frac{1}{2}(\eta_d - \mu)^\top \Sigma^{-1}(\eta_d - \mu))}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}},$$

где $|T|$ -мерный вектор μ и ковариационная $|T| \times |T|$ -матрица Σ являются параметрами многомерного нормального распределения. Векторы документов $\eta_d \in \mathbb{R}^{|T|}$ определяются через θ_d с точностью до константы C_d , зависящей от документа:

$$\eta_{td} = \ln \theta_{td} + C_d.$$

Изначально модель СТМ была разработана в рамках байесовского подхода, несмотря на дополнительные технические трудности, которые пришлось преодолевать из-за того, что лог-нормальное распределение не является сопряжённым к мультиномиальному. Мы покажем, что сама идея СТМ может быть гораздо проще реализована и понятнее изложена в терминах не-байесовской регуляризации.

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

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

Согласно (11), формула М-шага для θ_{td} принимает вид

$$\theta_{td} \propto \left(n_{td} - \tau \sum_{s \in T} \tilde{\Sigma}_{ts} (\ln \theta_{sd} - \mu_s) \right)_+, \quad (14)$$

где $\Sigma^{-1} = (\tilde{\Sigma}_{ts})_{T \times T}$ — обратная ковариационная матрица.

Предполагается, что параметры Σ, μ нормального распределения не изменяются во время одной итерации. Следуя идее блочно-покоординатной оптимизации, будем оценивать эти параметры модели по окончании каждого прохода коллекции, в Алгоритме 2.1 — после шага 9:

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

Таким образом, трудоёмкая операция обращения ковариационной матрицы выполняется только в конце каждой итерации. После обращения в матрице $\tilde{\Sigma}_{ts}$ можно обнулить незначимые элементы, чтобы, благодаря её разреженности, сократить вычисления по формуле (14). В [24] предлагалось использовать LASSO-регрессию, чтобы сразу получить разреженную ковариационную матрицу.

Максимизация когерентности. Тема называется *когерентной*, если термины, наиболее частые в данной теме, неслучайно часто совместно встречаются рядом в документах коллекции [25, 26]. Когерентность может оцениваться как по самой коллекции D [27], так и по сторонней коллекции, например, по Википедии [28]. Средняя когерентность тем считается хорошей мерой интерпретируемости тематической модели [26].

Пусть заданы оценки совместной встречаемости $C_{wv} = \hat{p}(w | v)$ для пар терминов $(w, v) \in W^2$. Обычно C_{wv} оценивают как долю документов, содержащих термин v , в которых термин w встречается не далее чем через 10 слов от v .

Запишем по формуле полной вероятности условную вероятность $\hat{p}(w|t)$ через условные вероятности $\phi_{vt} = p(v|t)$ всех терминов v , когерентных с w :

$$\hat{p}(w|t) = \sum_{v \in W \setminus w} C_{wv} \phi_{vt} = \sum_{v \in W \setminus w} \frac{C_{wv} n_{vt}}{n_t}.$$

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

$$R(\Phi) = \tau \sum_{t \in T} n_t \sum_{w \in W} \hat{p}(w|t) \ln \phi_{wt} \rightarrow \max.$$

Формула М-шага, согласно (10), принимает вид

$$\phi_{wt} \propto n_{wt} + \tau \sum_{v \in W \setminus w} C_{wv} n_{vt}.$$

Эта же формула предлагалась в [27] для модели LDA и алгоритма сэмплирования Гиббса, с более сложным обоснованием через обобщённую урновую схему Пойя, и более сложной эвристической оценкой C_{wv} .

В работе [28] предлагалось использовать другой регуляризатор:

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

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

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

Регуляризатор для задач классификации. Пусть каждый документ d относится к подмножеству C_d конечного множества классов C . Задача заключается в том, чтобы выявить связи между классами и темами, улучшить качество тематической модели благодаря информации о классах и построить алгоритм классификации новых документов. Обычные методы классификации показывают неудовлетворительные результаты на больших текстовых коллекциях с большим числом несбалансированных, пересекающихся, взаимозависимых классов. Преимущество тематических моделей в том, что они учитывают все классы одновременно [29].

В роли классов могут выступать категории [29, 30], авторы [31], моменты времени [32, 17], цитируемые документы [23], цитируемые авторы [33], пользователи документов [34]. Для этих и других случаев разработано множество специальных моделей, см. обзоры [1, 29]. Все эти задачи легко формализуются в рамках ARTM.

Расширим вероятностное пространство до множества $D \times W \times T \times C$. Будем считать, что с каждым словом w в каждом документе d связана не только тема $t \in T$, но и класс $c \in C$. Примем *гипотезу условной независимости*, полагая, что класс документа определяется только его тематикой: $p(c|t) = p(c|d, t)$. Следуя

тематической модели Dependency LDA [29], определим распределение классов документов $p(c|d)$ через распределения классов тем $\psi_{ct} = p(c|t)$ и тем документов $\theta_{td} = p(t|d)$ по аналогии с основной тематической моделью (1):

$$p(c|d) = \sum_{t \in T} \psi_{ct} \theta_{td}, \quad (15)$$

где матрица классов тем $\Psi = (\psi_{ct})_{C \times T}$ является новым параметром модели.

Введём ещё одну гипотезу, теперь об *условной независимости* терминов и классов в каждом документе: $p(w, c|d) = p(w|d)p(c|d)$. Благодаря этому предположению логарифм правдоподобия распадается на логарифм правдоподобия модели PLSA $L(\Psi, \Theta)$ и регуляризатор $Q(\Psi, \Theta)$:

$$\ln \prod_{d \in D} \prod_{w \in d} p(w, c|d)^{n_{dw}} = L(\Phi, \Theta) + \tau Q(\Psi, \Theta) \rightarrow \max_{\Phi, \Theta, \Psi}; \quad (16)$$

$$Q(\Psi, \Theta) = \sum_{d \in D} \sum_{c \in C} m_{dc} \ln \sum_{t \in T} \psi_{ct} \theta_{td}, \quad (17)$$

где m_{dc} — эмпирические частоты классов, которые естественно задавать как $m_{dc} = n_d[c \in C_d]/|C_d|$, исходя из интерпретации вероятностного пространства. Коэффициент регуляризации τ можно положить равным 1. Варьируя его, можно сбалансировать требования точности модели содержимого документов $p(w|d)$ и модели классификации $p(c|d)$. Регуляризатор Q минимизирует взвешенную сумму дивергенций между модельными распределениями $p(c|d)$ и соответствующими эмпирическими частотами классов в документах $\frac{m_{dc}}{n_d}$:

В [29] предлагался довольно громоздкий вывод алгоритма сэмплирования Гиббса в рамках байесовского подхода. Следующая теорема показывает, что в ARTM задача решается по-прежнему с помощью ЕМ-подобного алгоритма.

Теорема 3. Если функция $R(\Phi, \Psi, \Theta)$ стохастических матриц Φ, Ψ, Θ непрерывно дифференцируема и (Φ, Ψ, Θ) является точкой локального максимума $L(\Phi, \Theta) + \tau Q(\Psi, \Theta) + R(\Phi, \Psi, \Theta)$, то для любой регулярной темы t и любого регулярного документа d выполняется система уравнений:

$$p_{tdw} = \frac{\phi_{wt} \theta_{td}}{\sum_{s \in T} \phi_{ws} \theta_{sd}}; \quad p_{tdc} = \frac{\psi_{ct} \theta_{td}}{\sum_{s \in T} \psi_{cs} \theta_{sd}}; \quad (18)$$

$$\phi_{wt} \propto \left(n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right)_+; \quad n_{wt} = \sum_{d \in D} n_{dw} p_{tdw}; \quad (19)$$

$$\psi_{ct} \propto \left(m_{ct} + \psi_{ct} \frac{\partial R}{\partial \psi_{ct}} \right)_+; \quad m_{ct} = \sum_{d \in D} m_{dc} p_{tdc}; \quad (20)$$

$$\theta_{td} \propto \left(n_{td} + \tau m_{td} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right)_+; \quad n_{td} = \sum_{w \in d} n_{dw} p_{tdw}; \quad m_{td} = \sum_{c \in C_d} m_{dc} p_{tdc}. \quad (21)$$

Доказательство аналогично доказательству Теоремы 2.

Регуляризатор R может включать сглаживание матрицы Ψ , следуя примеру Dependency LDA, однако более естественным представляется разреживание. Другой пример регуляризации матрицы Ψ приводится далее.

Регуляризатор для балансирования классов (label regularization) [35, 29] улучшает качество классификации в задачах с большим числом несбалансированных, пересекающихся, взаимозависимых классов. Потребуем, чтобы оценка распределения классов $p(c)$, сделанная по тематической модели, была близка к частотам классов \hat{p}_c в выборке документов:

$$R(\Psi) = \xi \sum_{c \in C} \hat{p}_c \ln p(c) \rightarrow \max, \quad p(c) = \sum_{t \in T} \psi_{ct} p(t), \quad p(t) = \frac{n_t}{n}.$$

Отсюда следует формула М-шага:

$$\psi_{ct} \propto m_{ct} + \xi \hat{p}_c \frac{\psi_{ct} n_t}{\sum_{s \in T} \psi_{cs} n_s},$$

в которой распределения ψ_{ct} сглаживаются пропорционально частотам классов \hat{p}_c .

5. Комбинирование регуляризаторов для разреживания и улучшения интерпретируемости тем

Интерпретируемость тематической модели является плохо формализуемым требованием. Содержательно оно означает, что по спискам наиболее частотных слов и документов темы эксперт может понять, о чём эта тема, и дать ей адекватное название. Свойство интерпретируемости важно в информационно-поисковых системах для систематизации и визуализации результатов тематического поиска или категоризации документов.

Большинство существующих методов оценивания интерпретируемости основано на привлечении экспертов-ассессоров. В [36] экспертам предлагалось непосредственно оценивать полезность тем по трёхбалльной шкале. В методе интрузий [37] для каждой найденной темы составляется список из 10 наиболее частотных слов, в который внедряется одно случайное слово. Тема считается интерпретируемой, если подавляющее большинство экспертов правильно указывают лишнее слово. Экспертные подходы необходимы на стадии исследований, но они затрудняют автоматическое построение тематических моделей. В серии работ [36, 25, 25, 27] удалось найти величину, которая вычисляется по коллекции автоматически и хорошо коррелирует с экспертными оценками интерпретируемости. Это когерентность (coherence), оценивающая, насколько часто наиболее вероятные слова темы встречаются рядом в документах данной коллекции или во внешней политематической коллекции, такой, как Википедия. Когерентность на сегодняшний день остается основной мерой интерпретируемости, вычисляемой автоматически.

В данной работе предлагается другой подход к формализации понятия интерпретируемости и вводятся дополнительные меры интерпретируемости, также не требующие привлечения ассессоров. Предполагается, что интерпретируемая тема должна содержать лексическое ядро — множество слов, характерных для определённой предметной области, которые часто употребляются рядом в документах, с большой вероятностью употребляются в данной теме и практически не употребляются в других темах. Отсюда следует, что из бесконечного множества стохастических матричных разложений $F \approx \Phi \Theta$ нас больше всего интересуют те, в которых матрицы Φ и Θ обладают структурой разреженности, примерно показанной на рис. 3. Множество тем разбивается на два подмножества, $T = S \sqcup B$: предметные темы S и фоновые темы B .

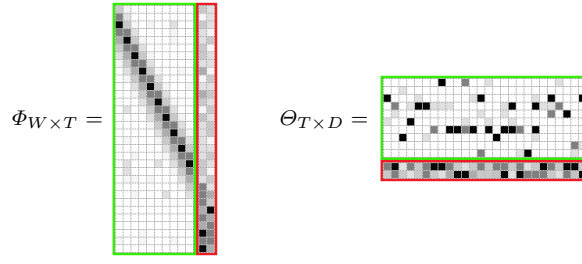


Рис. 3. Структура разреженности матриц Φ и Θ с предметными и фоновыми темами.

Предметные темы $t \in S$ содержат термины предметных областей. Их распределения $p(w|t)$ разрежены и существенно различны (декоррелированы). Распределения $p(d|t)$ также разрежены, так как каждая предметная тема присутствует в относительно небольшой доле документов.

Фоновые темы $t \in B$ содержат слова общей лексики, которых не должно быть в предметных темах. Их распределения $p(w|t)$ и $p(d|t)$ сглажены, так как эти слова присутствуют в большинстве документов. Тематическую модель с фоновыми темами можно рассматривать как обобщение робастных моделей [38, 15], в которых использовалось только одно фоновое распределение.

Комбинирование регуляризаторов. Для обеспечения структуры разреженности матриц Φ и Θ , показанной на рис. 3, предлагается комбинация из пяти регуляризаторов: сглаживание фоновых тем в матрицах Φ и Θ , разреживание предметных тем в матрицах Φ и Θ , и декоррелирование предметных тем в матрице Φ :

$$\begin{aligned} R(\Phi, \Theta) = & -\beta_0 \sum_{t \in S} \sum_{w \in W} \beta_w \ln \phi_{wt} - \alpha_0 \sum_{d \in D} \sum_{t \in S} \alpha_t \ln \theta_{td} \\ & + \beta_1 \sum_{t \in B} \sum_{w \in W} \beta_w \ln \phi_{wt} + \alpha_1 \sum_{d \in D} \sum_{t \in B} \alpha_t \ln \theta_{td} \\ & - \gamma \sum_{t \in T} \sum_{s \in T \setminus t} \sum_{w \in W} \phi_{wt} \phi_{ws} \rightarrow \max. \end{aligned}$$

где в качестве фоновое распределения β можно брать либо равномерное распределение, либо частоты слов в коллекции $\beta_w = n_w/n$; в качестве α естественно использовать равномерное распределение.

Формулы М-шага для комбинированной модели выписываются согласно (??):

$$\begin{aligned} \phi_{wt} \propto & \left(n_{wt} - \underbrace{\beta_0 \beta_w [t \in S]}_{\text{разреживание предметных тем}} + \underbrace{\beta_1 \beta_w [t \in B]}_{\text{сглаживание фоновых тем}} - \underbrace{\gamma [t \in S] \phi_{wt} \sum_{s \in S \setminus t} \phi_{ws}}_{\text{декоррелирование}} \right)_+; \\ \theta_{td} \propto & \left(n_{td} - \underbrace{\alpha_0 \alpha_t [t \in S]}_{\text{разреживание предметных тем}} + \underbrace{\alpha_1 \alpha_t [t \in B]}_{\text{сглаживание фоновых тем}} \right)_+. \end{aligned}$$

Траектории регуляризации. При линейном комбинировании регуляризаторов R_i возникает проблема выбора вектора коэффициентов $\tau = (\tau_i)_{i=1}^r$. Эффективный способ их оптимизации применяется в эластичных сетях (elastic net) для задач регрессии и классификации [39], однако он подходит только для комбинирования L_1 и L_2 -регуляризаторов. В тематическом моделировании разнообразие регуляризаторов гораздо больше. При чрезмерно больших значениях коэффициентов некоторые регуляризаторы могут конфликтовать друг с другом, ухудшать сходимость вдали от множества решений или приводить к вырождению модели. С другой стороны, при чрезмерно низких значениях коэффициентов регуляризаторы могут плохо выполнять свои функции. В теории решения некорректно поставленных обратных задач [3] известно, что для достижения множества решений коэффициенты регуляризации должны в ходе итераций сходиться к нулю. Однако оптимальный темп этой сходимости существенно зависит от конкретной задачи, и, как правило, его приходится подбирать экспериментально.

Будем называть *траекторией регуляризатора* функцию его коэффициента регуляризации от номера итерации и критериев качества модели. Будем подбирать траектории регуляризаторов экспериментальным путём, анализируя их влияние на критерии качества модели в ходе итераций.

Измерение качества модели. Поскольку задача построения тематической модели является многокритериальной, то и измерение качества модели должно вестись по совокупности критериев. Не претендуя на полноту, перечислим критерии, которые мы использовали в наших экспериментах.

Точность тематической модели $p(w|d)$ на коллекции D принято измерять с помощью *перплексии*, которая тесно связана с правдоподобием (чем ниже перплексия, тем лучше):

$$\mathcal{P}(D, p) = \exp\left(-\frac{1}{n}L(\Phi, \Theta)\right) = \exp\left(-\frac{1}{n} \sum_{d \in D} \sum_{w \in d} n_{dw} \ln p(w|d)\right). \quad (22)$$

Контрольная перплексия $\mathcal{P}(D', p_D)$ 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. Если во второй половине оказывались термины, которых не было в обучающей коллекции D , то они игнорировались. Параметры ϕ_t оценивались только по обучающей коллекции.

Разреженность модели измерялась долей \mathcal{S}_Φ и \mathcal{S}_Θ нулевых элементов, соответствующих предметным темам в матрицах Φ и Θ ,

Доля фоновых слов во всей коллекции

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

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

Интерпретируемость тематической модели оценивалась несколькими критериями, характеризующими близость матрицы Φ к разреженной структуре, показанной на рис. 3. Определим ядро W_t темы t как множество терминов, которые имеют высокую условную вероятность $p(t|w) = \phi_{wt} \frac{n_t}{n_w}$ для данной темы:

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

По ядру определим три показателя интерпретируемости темы t :

$\text{pur}_t = \sum_{w \in W_t} p(w|t)$ — чистота темы (чем выше, тем лучше);

$\text{con}_t = \frac{1}{|W_t|} \sum_{w \in W_t} p(t|w)$ — контрастность темы (чем выше, тем лучше);

$\text{ker}_t = |W_t|$ — размер ядра (ориентировочный оптимум $\frac{|W|}{|T|}$).

Когерентность темы t измерялась как средняя *поточечная взаимная информация* по всем парам k наиболее вероятных слов темы t [25]:

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

где w_i — i -й термин в порядке убывания ϕ_{wt} . 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.

Показатели когерентности, размера ядра, чистоты и контрастности модели определим как средние по всем предметным темам $t \in S$.

Исходные данные. Эксперименты проводились на коллекции NIPS, которая содержит $|D| = 1566$ текстов статей научной конференции Neural Information Processing Systems на английском языке. Суммарная длина коллекции $n \approx 2.3 \cdot 10^6$ слов. Объём словаря $|W| \approx 1.3 \cdot 10^4$. Контрольная коллекция D' содержит 174 документа. Предварительная обработка текстов включала приведение к нижнему регистру, удаление пунктуации, удаление стоп-слов с помощью библиотеки BOW toolkit [40].

Результаты экспериментов. Во всех экспериментах фиксировалось число тем $|T| = 100$, из них фоновых тем $|B| = 10$, число итераций 40.

В таблице 1 приводятся результаты сравнения тематических моделей. Первые две строки соответствуют стандартным моделям PLSA и LDA, остальные строки — регуляризованным моделям ARTM. Первые три колонки задают комбинации регуляризаторов сглаживания, разреживания и декоррелирования. Остальные колонки соответствуют введённым выше критериям качества.

Для оценивания модели LDA использовался регуляризованный EM-алгоритм с параметрами сглаживания $\alpha = 0.5$, $\beta = 0.01$, соответствующими симметричному распределению Дирихле.

Для сглаживания фоновых тем использовались равномерные распределения при коэффициентах регуляризации $\alpha = 0.8$, $\beta = 0.1$.

Для разреживания предметных тем в столбцах матрицы Φ использовалось одно из двух распределений: равномерное $\beta_w = \frac{1}{|W|}$ или фоновое $\beta_w = \frac{n_w}{n}$.

Основной вывод заключается в том, что комбинирование регуляризаторов позволяет улучшить все критерии качества при незначительном ухудшении перплексии. Разреживание обнуляет до 96% элементов матрицы Φ и до 87% элементов

Таблица 1. Сравнение регуляризованных тематических моделей со сглаживанием (Sm), разреживанием (Sp) по равномерному (u) или фоновому (b) распределению и декоррелированием (Dc). Критерии: \mathcal{P} — контрольная перплексия, \mathcal{B} — доля фоновых слов в коллекции, \mathcal{S}_Φ и \mathcal{S}_Θ — разреженность матриц Φ и Θ , con — контрастность, pur — чистота, ker — размер ядра, \mathcal{C}^{ker} — когерентность ядра, \mathcal{C}^{10} и \mathcal{C}^{100} — когерентность 10 и 100 наиболее вероятных слов. Выделены лучшие значения в каждой колонке.

Sm	Sp	Dc	\mathcal{P}	\mathcal{B}	\mathcal{S}_Φ	\mathcal{S}_Θ	con	pur	ker	\mathcal{C}^{ker}	\mathcal{C}^{10}	\mathcal{C}^{100}
—	—	—	1923	0.00	0.000	0.000	0.43	0.14	100	0.84	0.25	0.17
+	—	—	1902	0.00	0.000	0.000	0.42	0.12	82	0.93	0.26	0.17
—	u	—	2114	0.24	0.957	0.867	0.53	0.20	71	0.91	0.25	0.18
—	b	—	2507	0.51	0.957	0.867	0.46	0.56	151	0.71	0.60	0.58
—	—	+	2025	0.57	0.561	0.000	0.46	0.38	109	0.82	0.94	0.56
+	u	—	1961	0.25	0.957	0.867	0.51	0.20	64	0.97	0.26	0.18
+	b	—	2025	0.49	0.957	0.867	0.45	0.52	128	0.77	0.55	0.55
+	—	+	1985	0.59	0.582	0.000	0.46	0.39	97	0.87	0.93	0.57
+	u	+	2010	0.73	0.980	0.867	0.56	0.73	78	0.94	0.94	0.62
+	b	+	2026	0.80	0.979	0.867	0.52	0.89	111	0.81	0.96	0.83

матрицы Θ . Декорреляция повышает чистоту и когерентность тем. Сглаживание фоновых тем помогает им очистить предметные темы от слов общей лексики. Все эти улучшения сопровождаются незначительной потерей перплексии, что согласуется с наблюдениями и выводами из [37] о том, что модели, имеющие лучшую перплексию, то есть лучше предсказывающие появление слов в документах, часто демонстрируют худшую интерпретируемость латентных тем.

Для мониторинга процесса построения модели и подбора траекторий регуляризации строились графики зависимости критериев качества модели от номера итерации. На рис. 4, 5, 6 критерии качества откладываются по вертикальным осям: на левом графике по левой оси — контрольная перплексия \mathcal{P} , по правой оси — разреженности матриц \mathcal{S}_Φ и \mathcal{S}_Θ и доля фоновых слов \mathcal{B} ; на среднем графике по левой оси — размер ядра ker, по правой оси — контрастность con и чистота pur; на правом графике по левой оси — когерентности \mathcal{C}^{ker} , \mathcal{C}^{10} и \mathcal{C}^{100} . Такие графики дают понимание эффектов каждого регуляризатора в отдельности и в комбинации с остальными.

Рис. 4 показывает, что модель PLSA не разреживает матрицы Φ и Θ , не позволяет определять фоновые слова и даёт низкую чистоту тем.

Рис. 5 показывает совокупное влияние разреживания предметных тем (по фоновому распределению β_w) и сглаживания фоновых тем.

Рис. 6 позволяет увидеть дополнительные эффекты декоррелирования. Декоррелирование увеличивает чистоту и когерентность тем, очищает темы от слов общей лексики, при этом доля фоновых слов во всей коллекции достигает почти 80%.

Интересно отметить, что критерии качества могут существенно изменяться после достижения сходимости правдоподобия модели, то есть при неизменной перплексии или при незначительном её ухудшении.

Ввиду ограничений объёма мы не можем показать все варианты сравнений, которые были сделаны в экспериментах, и приводим только окончательные рекомендации по выбору траекторий регуляризации.

Коэффициенты регуляризации для разреживания предметных тем рекомендуется включить только после того, как итерационный процесс начал сходиться и определились близкие к нулю элементы матриц Φ и Θ . Более раннее или более резкое разреживание может ухудшать перплексию. Мы включали разреживание,

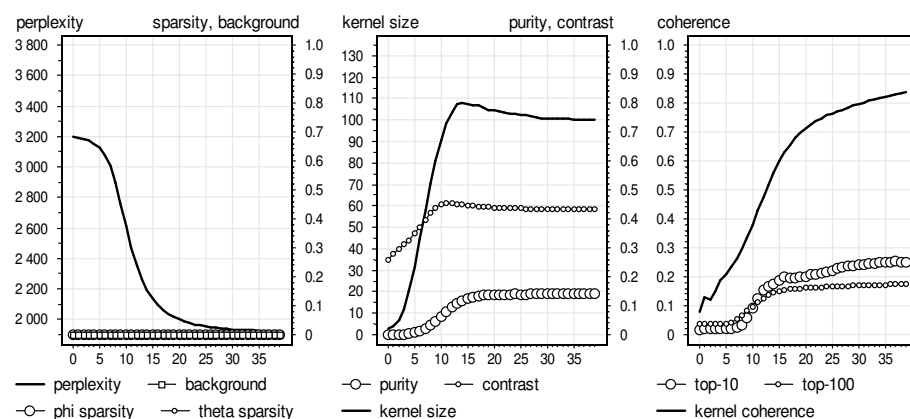


Рис. 4. Мониторинг критериев качества модели PLSA.

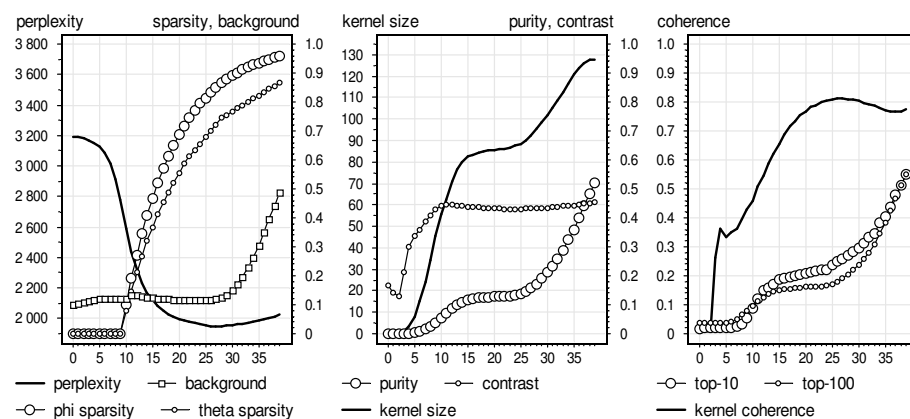


Рис. 5. Мониторинг модели ARTM со сглаживанием и разреживанием.

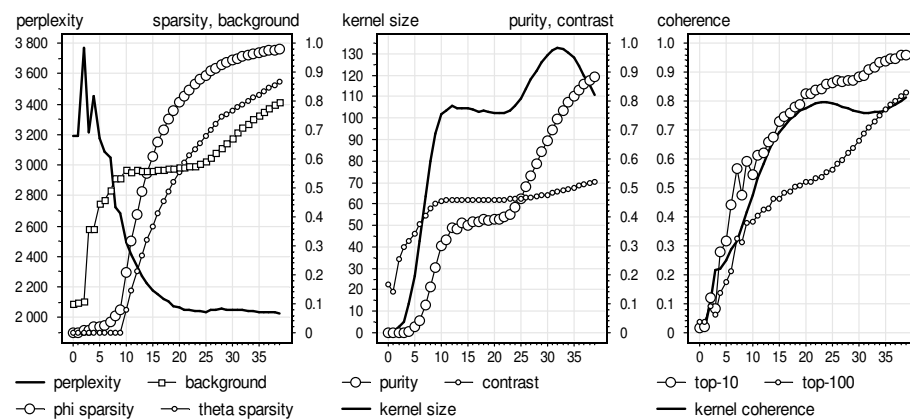


Рис. 6. Мониторинг модели ARTM со сглаживанием, разреживанием и декорреляцией.

начиная с 10-й итерации, обнуляя на каждой итерации 8% ненулевых значений в каждом векторе θ_d и 10% в каждом векторе ϕ_t .

Декоррелирование предметных тем включалось с первой итерации, коэффициент регуляризации был выбран постоянным и наибольшим, при котором ещё не происходило существенного увеличения перплексии, для данной коллекции было подобрано значение $\gamma = 2 \cdot 10^5$.

Сглаживание фоновых тем также оказалось лучше включать с первой итерации, не меняя коэффициент регуляризации в ходе итераций.

6. Обсуждение и выводы

Построение тематической модели текстовой коллекции является некорректно поставленной задачей стохастического матричного разложения. Множество её решений в общем случае бесконечно, поэтому численные решения неустойчивы и зависят от случайного начального приближения. Байесовская регуляризация в методе латентного размещения Дирихле (LDA) не решает данную проблему, поскольку распределение Дирихле является слишком слабым регуляризатором. Для повышения устойчивости необходимы проблемно-ориентированные регуляризаторы, формализующие более сильные предположения о структуре матричного разложения.

В данной работе предлагается полу-вероятностный подход к моделированию тематики текстовых коллекций — *аддитивная регуляризация тематических моделей* (ARTM). Он основан на максимизации взвешенной суммы критериев регуляризации. Построение тематической модели рассматривается как задача многокритериальной оптимизации, которая сводится к однокритериальной задаче путём скаляризации критериев.

Для решения оптимизационной задачи предлагается регуляризованный ЕМ-алгоритм, в который можно подставлять любые регуляризаторы или их линейные комбинации. По сравнению с доминирующим байесовским подходом, ARTM позволяет отказаться от избыточных вероятностных допущений, упростить математический аппарат и использовать произвольные сочетания регуляризаторов.

ARTM приводит к модульной технологии тематического моделирования, основанной на библиотеке регуляризаторов с унифицированным интерфейсом. Для построения тематической модели в прикладной задаче достаточно выбрать комбинацию регуляризаторов, покрывающих требования данной задачи.

В данной работе мы рассмотрели общий подход ARTM при некоторых ограничениях, снятие которых будет предметом ближайших исследований.

Во-первых, мы ограничились моделями, основанными на гипотезе «мешка слов». Не рассматривались тематические модели с более сложной структурой — иерархические, мультиграммные, мультязычные и другие, в которых применение регуляризации, возможно, потребует дополнительных усилий.

Во-вторых, мы рассмотрели только один численный метод — ЕМ-алгоритм, позволяющий включать в модель широкий класс регуляризаторов. Альтернативные способы решения возникающих систем уравнений, а также вопросы сходимости и устойчивости решений пока не рассматривались.

В третьих, мы привели далеко не полный обзор регуляризаторов, а в экспериментальной части работы ограничились только тремя регуляризаторами сглаживания, разреживания и декоррелирования. Они повышают интерпретируемость

тем и образуют минимальный необходимый набор регуляризаторов для большинства задач тематического моделирования. Масштабные эксперименты с комбинациями из большого числа регуляризаторов выходят за рамки данной работы.

Наконец, столкнувшись с проблемой оптимизации траектории в пространстве коэффициентов регуляризации, мы ограничились методикой визуального сравнения и мониторинга качества моделей в процесс итераций и выработкой эмпирических рекомендаций для выбранной комбинации из трёх регуляризаторов.

Работа выполнена при поддержке Российского фонда фундаментальных исследований (проекты 14-07-00847, 14-07-00908) и программы ОМН РАН «Алгебраические и комбинаторные методы математической кибернетики и информационные системы нового поколения».

Авторы признательны А. И. Фрею и М. Н. Рыскиной за полезные обсуждения и помощь с переводом статьи на английский язык, а также В. В. Глушаченкову за предоставленные результаты экспериментов на модельных данных.

Список литературы

1. 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.
2. Blei D. M., Ng A. Y., Jordan M. I. Latent Dirichlet allocation // *Journal of Machine Learning Research*. — 2003. — Vol. 3. — Pp. 993–1022.
3. Tikhonov A. N., Arsenin V. Y. Solution of ill-posed problems. — W. H. Winston, Washington, DC, 1977.
4. Khalifa O., Corne D., Chantler M., Halley F. Multi-objective topic modelling // 7th International Conference Evolutionary Multi-Criterion Optimization (EMO 2013). — Springer LNCS, 2013. — Pp. 51–65.
5. Si L., Jin R. Adjusting mixture weights of gaussian mixture model via regularized probabilistic latent semantic analysis // Proceedings of the Ninth Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD) / Ed. by T. B. Ho, D. W.-L. Cheung, H. Liu. — Vol. 3518 of *Lecture Notes in Computer Science*. — Springer, 2005. — Pp. 622–631.
6. Chien J.-T., Wu M.-S. Adaptive bayesian latent semantic analysis // *IEEE Transactions on Audio, Speech, and Language Processing*. — 2008. — Vol. 16, no. 1. — Pp. 198–207.
7. Wang Q., Xu J., Li H., Craswell N. Regularized latent semantic indexing // SIGIR. — 2011. — Pp. 685–694.
8. 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.
9. 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.
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. — Pp. 27–34.
11. Wang Y. Distributed Gibbs sampling of latent dirichlet allocation: The gritty details. — 2008.
12. 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.
13. 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.

14. Lu Y., Mei Q., Zhai C. Investigating task performance of probabilistic topic models: an empirical study of PLSA and LDA // *Information Retrieval*. — 2011. — Vol. 14, no. 2. — Pp. 178–203.
15. 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.
16. 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.
17. 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.
18. Shashanka M., Raj B., Smaragdis P. Sparse overcomplete latent variable decomposition of counts data // *Advances in Neural Information Processing Systems, NIPS-2007* / Ed. by J. C. Platt, D. Koller, Y. Singer, S. Roweis. — Cambridge, MA: MIT Press, 2008. — Pp. 1313–1320.
19. Eisenstein J., Ahmed A., Xing E. P. Sparse additive generative models of text // *ICML'11*. — 2011. — Pp. 1041–1048.
20. Wang C., Blei D. M. Decoupling sparsity and smoothness in the discrete hierarchical dirichlet process // *NIPS*. — Curran Associates, Inc., 2009. — Pp. 1982–1989.
21. Chien J.-T., Chang Y.-L. Bayesian sparse topic model // *Journal of Signal Processing Systems*. — 2013. — Pp. 1–15.
22. Tan Y., Ou Z. Topic-weak-correlated latent dirichlet allocation // 7th International Symposium Chinese Spoken Language Processing (ISCSLP). — 2010. — Pp. 224–228.
23. Dietz L., Bickel S., Scheffer T. Unsupervised prediction of citation influences // *Proceedings of the 24th international conference on Machine learning*. — ICML '07. — New York, NY, USA: ACM, 2007. — Pp. 233–240.
24. Blei D., Lafferty J. A correlated topic model of Science // *Annals of Applied Statistics*. — 2007. — Vol. 1. — Pp. 17–35.
25. 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.
26. 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.
27. 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.
28. 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.
29. Rubin T. N., Chambers A., Smyth P., Steyvers M. Statistical topic models for multi-label document classification // *Machine Learning*. — 2012. — Vol. 88, no. 1-2. — Pp. 157–208.
30. Zhou S., Li K., Liu Y. Text categorization based on topic model // *International Journal of Computational Intelligence Systems*. — 2009. — Vol. 2, no. 4. — Pp. 398–409.
31. Rosen-Zvi M., Griffiths T., Steyvers M., Smyth P. The author-topic model for authors and documents // *Proceedings of the 20th conference on Uncertainty in artificial intelligence*. — UAI '04. — Arlington, Virginia, United States: AUAI Press, 2004. — Pp. 487–494.
32. TextFlow: Towards better understanding of evolving topics in text. / W. Cui, S. Liu, L. Tan, C. Shi, Y. Song, Z. Gao, H. Qu, X. Tong // *IEEE transactions on visualization and computer graphics*. — 2011. — Vol. 17, no. 12. — Pp. 2412–2421.
33. Kataria S., Mitra P., Caragea C., Giles C. L. Context sensitive topic models for author influence in document networks // *Proceedings of the Twenty-Second international joint conference on Artificial Intelligence* — Volume 3. — IJCAI'11. — AAAI Press, 2011. — Pp. 2274–2280.
34. Wang C., Blei D. M. Collaborative topic modeling for recommending scientific articles // *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. — New York, NY, USA: ACM, 2011. — Pp. 448–456.

35. Mann G. S., McCallum A. Simple, robust, scalable semi-supervised learning via expectation regularization // Proceedings of the 24th international conference on Machine learning. — ICML '07. — New York, NY, USA: ACM, 2007. — Pp. 593–600.
36. Newman D., Karimi S., Cavedon L. External evaluation of topic models // Australasian Document Computing Symposium. — December 2009. — Pp. 11–18.
37. Chang J., Gerrish S., Wang C., Boyd-Graber J. L., Blei D. M. Reading tea leaves: How humans interpret topic models // Neural Information Processing Systems (NIPS). — 2009. — Pp. 288–296.
38. Chemudugunta C., Smyth P., Steyvers M. Modeling general and specific aspects of documents with a probabilistic topic model // Advances in Neural Information Processing Systems. — MIT Press, 2007. — Vol. 19. — Pp. 241–248.
39. 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.
40. McCallum A. K. Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. — <http://www.cs.cmu.edu/~mccallum/bow>.