Московский Физико-Технический Институт
(Государственный Университет)

Факультет Управления и Прикладной Математики
Кафедра «Интеллектуальные Системы»

ДИПЛОМНАЯ РАБОТА СТУДЕНТА 174 ГРУППЫ

«Отбор тем в задачах тематического моделирования»

Выполнил:
студент 4 курса 174 группы
Плавин Александр Викторович

Научный руководитель:
проф. каф. Интеллектуальные системы, д. ф.-м. н.
Воронцов Константин Вячеславович

Москва, 2015

Содержание

Аннотация

В работе строится алгоритм порождения ранжирующих функций для задачи информационного поиска. Ранжирующие функции ищутся в виде суперпозиции заданных порождающих функций и переменных. Используется генетический алгоритм порождения суперпозиций. Структурная сложность получаемых моделей контролируется регуляризатором. С помощью метрики на множестве моделей определяется момент попадания в локальный минимум. Итоговые функции сравниваются относительно функционала MAP со структурно-простыми суперпозициями, полученными алгоритмом полного перебора. Вычислительный эксперимент проводится на выборке, состоящей из коллекций документов Trec5-8, запросов к каждой из них и экспертных оценок релевантности.

Ключевые слова: тематическое моделирование, вероятностная модель, число тем, EM-алгоритм, регуляризация.

1 Введение

Тематическое моделирование — способ построения модели текстовой коллекции, которая определяет, к каким *темам* относится каждый из документов и какие слова образуют каждую тему. Это одно из быстро развивающихся направлений статистического анализа текстов, которое позволяет эффективно решать проблемы информационного поиска [1], кластеризации и классификации документов [2], построения рекомендательных систем [3], суммаризации больших коллекций. Тематические модели могут учитывать различные особенности языка и текстовых коллекций, например выделенные ключевые слова и фразы, структуру предложений и абзацев, изменение тематики во времени или внутри отдельных документов, различного рода связи между документами. В каждом случае могут ставиться различные дополнительные требования к модели, порождая множество модификаций [4].

Наибольшее применение в тематическом моделировании находят *вероятностные модели*. Они осуществляют «мягкую» кластеризацию, считая что документ или слово относится сразу к нескольким темам с различными вероятностями. Таким образом, каждая тема описывается вероятностным распределением на множестве слова, а каждый документ — распределением на множестве тем. Предполагается, что при выборе каждого слова в документе сначала выбирается тема из распределения для этого документа, а затем слово из распределения для этой темы (рис. 1). Построить вероятностную тематическую модель — значит решить задачу восстановления исходных распределений по известной коллекции.

Важной проблемой в построении тематической модели является определение числа тем, так как от него зависит качество получаемой модели и эффективность работы алгоритмов. Во многих подходах число тем является задаваемым извне параметром, который требуется найти заранее из некоторых внешних соображений. Существуют методы, определяющие число тем автоматически, но, как будет показано далее, самый популярный из них показывает неустойчивые результаты и неэффективен с вычислительной точки зрения.

Целью данной работы является разработка и исследование метода определения оптимального числа тем в вероятностных тематических моделях, основанного на по-

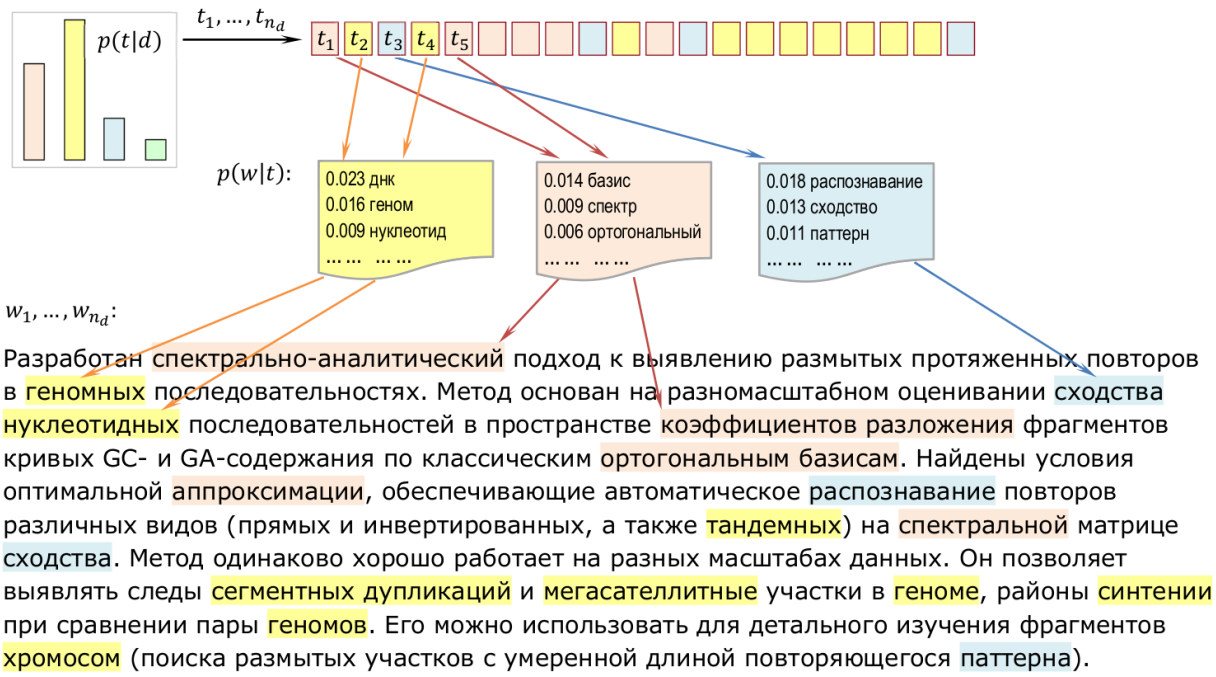


Рис. 1: Пример тематической структуры документа

степенном отборе тем. Поведение предлагаемого метода рассматривается как с теоретической точки зрения, так и путём вычислительных экспериментов с использованием реальных текстовых коллекций. Показывается, что таким способом действительно можно определять число тем, получаемые результаты устойчивы с нескольких точек зрения, а реализация метода вычислительно эффективна.

2 Постановка задачи

Пусть даны D — множество (*коллекция*) текстовых документов, W — множество (*словарь*) всех употребляемых в них терминов, которыми могут быть как отдельные слова, так и ключевые фразы. Каждый документ d — это последовательность n_d терминов $(w_1, \ldots, w_{n_d}) \subset W$, где термины могут повторяться, и каждому из них ставится в соответствие число n_{dw} его вхождений.

Согласно *гипотезе «мешка слов»*, порядок последовательности последовательности слов в каждом документе не важен и он не учитывается. Это позволяет значительно упростить выкладки и саму модель, хотя и потеряв некоторую часть важной для восприятия человеком информации: порядок слов в предложениях, предложений в тексте.

Предполагается, что существует конечное множество *тем* (скрытых переменных) T, и каждое употребление термина $w \in W$ в каждом документе $d \in D$ связано с некоторой неизвестной темой $t \in T$ (рис. 1). Считается, что верна *гипотеза условной независимости*: появление слов, относящихся к теме t, не зависит от документа, то есть соответствующее распределение является общим для всей коллекции:

$$p(w|d,t) = p(w|t).$$

Тогда, согласно формуле полной вероятности:

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

Эта гипотеза является формализацией предположения о том, что для написания каждого слова в документе сначала выбирается тема t из распределения $p(t|d)$, а затем само слово из распределения $p(w|t)$, не зависящего от документа.

Задача построения модели состоит в нахождении стохастических матриц *терминов тем* Φ и *тем документов* Θ, столбцы которых являются соответствующими вероятностными распределениями:

$$\Phi = (\phi_{wt})_{W \times T} = (p(w|t))_{W \times T},$$

$$\Theta = (\theta_{td})_{T \times D} = (p(t|d))_{T \times D}.$$

Нахождение таких матриц, которые порождают заданную коллекцию D с известными n_{dw}, представляет собой максимизацию некоторого критерия схожести реального и модельного распределений:

$$\frac{n_{dw}}{n_d} \text{ и } p(w|d).$$

В качестве такого критерия обычно используется *правдоподобие*:

$$\mathscr{L} = \prod_{d \in D, w \in d} p(w|d)^{n_{dw}} \to \max_{\Phi,\Theta} \tag{2}$$

или его логарифм $L = \ln \mathscr{L}$, для удобства вычислений.

3 Смежные исследования

Базовым способом построения тематических моделей является вероятностный латентный семантический анализ (PLSA [5]), на котором основано большинство современных методов. PLSA заключается в максимизации правдоподобия (2) с помощью итерационного процесса. Основным недостатком этого метода является неоднозначность решения в силу некорректной постановки задачи (2) [6, 7], и как следствие неустойчивость получаемых результатов относительно случайного начального приближения алгоритма.

Часто разрабатываются и используются модели на основе латентного размещения Дирихле (LDA [8]). В этой модели предполагается, что распределения слов в темах и тем в документах выбираются из распределений Дирихле с некоторыми параметрами, а обучение модели производится с помощью аппарата Байесовского вывода. Выбор именно распределения Дирихле в качестве априорного удобен с математической точки зрения, так как позволяет провести значительную часть вычислений аналитически. На LDA основано множество расширенных моделей, учитывающих различные дополнительные требования, однако их совмещение является сложной задачей и в каждом случае проводится отдельно.

Также недостатком является то, что как базовый PLSA, так и LDA требуют явного задания числа тем в качестве параметра алгоритма. Аналогичная проблема — определение числа кластеров — существует в любом алгоритме кластеризации, так как при построении модели априори неясно, какое значение брать. Правильное определение числа тем также требуется для получения наиболее интерпретируемой человеком модели: в случае значительного отклонения от истинного количества, темы коллекции могут смешиваться или наоборот, разбиваться на несколько и становиться незначимыми или очень близкими. Также выбор неоправданно большого числа тем может приводить к неэффективности алгоритма как по памяти, так и по времени работы.

Проблема оптимизации числа тем обычно решается с использованием расширения LDA — модели иерархических процессов Дирихле (HDP) [9]. Однако эта модель определяет число тем неустойчиво, и результат существенно зависит как от выбран-

ного начального приближения, так и от параметров алгоритма. Также остаются указанные выше трудности с совмещением различных требований в одной модели.

Другим вариантом решения проблемы выбора числа тем является непараметрический PLSA (nPLSA), предложенный в [10]. Отличие от PLSA заключается в том, что новые темы добавляются в процессе работы и их число не фиксировано. Последовательно выбирается документ, хуже всего описываемый существующими темами (с минимальным *правдоподобием*), и в модель добавляется тема с распределением слов этого документа. Для остальных документов при этом проводится *вклейка* (fold-in) этой темы. Такие шаги повторяются до достижения максимума *разнообразия* (diversity) тем, то есть среднего попарного расстояния между ними:

$$Diversity(\Phi) = \left\langle dist(\phi_{*t_i}, \phi_{*t_j}) \right\rangle_{i,j}.$$

После достижения максимума добавление тем прекращается и продолжается обучение обычной модели PLSA до сходимости.

Также в [10] предложен метод нахождения числа тем с использованием заданного пользователем поискового запроса, характеризующего одну из искомых тем. В таком случае описанный выше процесс останавливается по достижению максимальной близости одной из тем к запросу. Предполагается, что в этом случае одна из найденных тем с высокой точностью соответствует заданному поисковому запросу, а остальные темы имеют аналогичный размер и поэтому являются именно тем, что хотел получить пользователь.

Важным подходом, на базе которого предлагается проводить определение числа тем в данной работе, является подход *аддитивной регуляризации тематических моделей* (ARTM) [6], обобщающий PLSA. В нём используется то, что в общем случае конкретные решения некорректно поставленных задач можно находить с помощью *регуляризации* [11]. Регуляризация рассматриваемой задачи (2) возможна путём добавления дополнительных слагаемых (*регуляризаторов* R_i с некоторыми *коэффициентами регуляризации* τ_i) к максимизируемому критерию в задаче (2):

$$L + \sum_i \tau_i R_i \to \max_{\Phi,\Theta} \tag{3}$$

Такой подход предоставляет универсальный способ построения гибких многофункциональных моделей: как показано в [7, 12], ARTM позволяет находить разреженные,

хорошо интерпретируемые и когерентные темы, а также позволяет естественным образом обобщить модель для учёта метаданных в документах. В данной работе предлагается метод определения числа тем в коллекции в рамках этого подхода.

Алгоритм 1 EM-алгоритм обучения модели PLSA.

Вход: коллекция D, число тем $|T|$;

Выход: распределения Φ и Θ;

присвоить Φ и Θ случайные стохастические матрицы;

повторять

E-шаг: вспомогательные условные вероятности: $p(t|d,w) = \frac{p(w|t)p(t|d)}{p(w|d)}$;

M-шаг: новое приближение: $\hat{n}_{wt} = \sum_d n_{dw} p(t|d,w)$, $\hat{n}_{td} = \sum_w n_{dw} p(t|d,w)$;

получить стохастические матрицы: $\phi_{wt} = \frac{\hat{n}_{wt}}{\sum_{w'} \hat{n}_{w't}}$, $\theta_{td} = \frac{\hat{n}_{td}}{\sum_{t'} \hat{n}_{t'd}}$;

пока матрицы не сошлись;

4 Построение тематической модели ARTM

Как показано в (2), базовая модель PLSA для нахождения Φ и Θ использует максимизацию логарифма *правдоподобия*:

$$L = \ln \mathscr{L} = \ln \prod_{d \in D, w \in d} p(w|d)^{n_{dw}} \to \max_{\Phi,\Theta}.$$

Преобразовав произведение в сумму и раскрыв вероятность $p(w|d)$ согласно (1), получим итоговую задачу оптимизации:

$$L(\Phi,\Theta) = \sum_{d \in D, w \in d} n_{dw} \ln \sum_{t \in T} \phi_{wt}\theta_{td} \to \max_{\Phi,\Theta}. \tag{4}$$

Стандартным методом решения такой задачи является *EM-алгоритм* [13, 14], который состоит из двух последовательно выполняемых шагов: E *(expectation)* и M *(maximization)*: алгоритм 1.

Соответственно, обобщённая задача оптимизации с регуляризаторами (3) принимает вид:

$$\sum_{d \in D, w \in d} n_{dw} \ln \sum_{t \in T} \phi_{wt}\theta_{td} + R(\Phi,\Theta) \to \max_{\Phi,\Theta}, \quad \text{где } R(\Phi,\Theta) = \sum_i \tau_i R_i(\Phi,\Theta) \tag{5}$$

Эта задача также решается EM-алгоритмом 1, но формулы M-шага несколько модифицируются:

$$\begin{aligned} \phi_{wt} &= \operatorname{norm}\left(\hat{n}_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right)_+ \quad \text{вместо } \phi_{wt} = \operatorname{norm}(\hat{n}_{wt}), \\ \theta_{td} &= \operatorname{norm}\left(\hat{n}_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right)_+ \quad \text{вместо } \theta_{td} = \operatorname{norm}(\hat{n}_{td}). \end{aligned} \tag{6}$$

Здесь $(z)_+ = \max\{z, 0\}$ — положительная срезка, а $\text{norm}(a_{ij}) = \frac{a_{ij}}{\sum_{i'} a_{i'j}}$ — нормирование вероятностного распределения. Как показано в [14, 15], модифицированный таким образом алгоритм находит стационарную точку в задаче оптимизации с регуляризаторами (5).

5 Энтропийный регуляризатор отбора тем

В данной работе предлагается метод отбора тем, основанный на ARTM, который даёт возможность определять оптимальное число тем в модели, а также получать более интерпретируемые темы. Предложенный метод основан на использовании *энтропийного регуляризатора* [16, 17], постепенно удаляющего темы из модели. В связи с общностью ARTM [6, 7, 12], предлагаемый способ может быть использован в комбинации с многими другими регуляризаторами для совмещения особенностей моделей.

Предлагаемый здесь энтропийный регуляризатор состоит в максимизации KL-дивергенции между равномерным распределением вероятностей тем:

$$p_U(t) = \frac{1}{|T|},$$

и получаемым в модели распределением:

$$p(t) = \sum_d p(t|d)p(d) = \sum_d \theta_{td}\frac{n_d}{n}.$$

Таким образом, добавляемое слагаемое в задаче (3) имеет вид:

$$R_1(\Phi,\Theta) = KL\left(p_U \| p\right) = KL\left(\frac{1}{|T|}\middle\| \sum_d \theta_{td}\frac{n_d}{n}\right) \tag{7}$$

Как будет показано далее, такой регуляризатор разреживает (обнуляет) строки матрицы Θ целиком, и поэтому также называется *регуляризатором строкового разреживания* матрицы Θ. Разреживание строк соответствует полному исключению отдельных тем из модели от итерации к итерации, то есть отбору тем. Таким образом, для правильной работы этого регуляризатора требуется задавать заведомо избыточное начальное приближение для числа тем.

После раскрытия обозначений (7) примет следующий вид [17]:

$$R_1(\Phi,\Theta) = \sum_t \frac{1}{|T|} \ln \frac{\frac{1}{|T|}}{\sum_d \theta_{td}\frac{n_d}{n}}, \tag{8}$$

что, согласно (6), даёт для формул M-шага:

$$\phi_{wt} = \mathrm{norm}(\hat{n}_{wt}),$$

$$\theta_{td} = \text{norm}\left(\hat{n}_{td} - \tau \frac{n}{|T|}\frac{n_d}{\hat{n}_t}\theta_{td}\right)_+.$$

Простую интерпретацию этого регуляризатора как строкового разреживания матрицы Θ можно получить, если заменить θ_{td} в правой части формулы её несмещённой оценкой $\theta_{td} \approx \frac{\hat{n}_{td}}{n_d}$:

$$\theta_{td} = \text{norm}\left(\hat{n}_{td} - \tau \frac{n}{|T|}\frac{n_d}{\hat{n}_t}\frac{\hat{n}_{td}}{n_d}\right)_+ = \text{norm}\left[\hat{n}_{td}\left(1 - \tau \frac{n}{|T|}\frac{1}{\hat{n}_t}\right)_+\right]. \tag{9}$$

Как видно, в таком случае все элементы строки будут обнулены, если $\hat{n}_t$ достаточно малое. Также отсюда видно, что характерные значения коэффициента регуляризации τ находятся в диапазоне $[0, 1]$, т.к. $|T|\hat{n}_t \sim n$.

В последующих разделах исследуется поведение этого регуляризатора с точки зрения определения оптимального числа тем в коллекции, а также получения более интерпретируемых тем.

6 Анализ предлагаемого подхода

Для определения оптимального числа тем предлагается запустить описанный алгоритм для различных значений коэффициента регуляризации τ и рассмотреть зависимость найденного числа тем от τ. Ожидается, что эта зависимость будет иметь участок, на котором число тем близко к постоянному, и тогда это число будем называть оптимальным.

Это предположение основано на следующих соображениях, которые относятся к базовому методу без регуляризаторов:

Теорема 1. *Пусть:*

- *коллекция точно соответствует идеализированной вероятностной модели* (1)*, т.е. для некоторых* Φ, Θ *выполнено:*

$$n_{dw} = n_d \sum_{t \in T} \phi_{wt} \theta_{td}.$$

- *некоторым образом эта коллекция идеально восстановлена в виде разложения на некоторые (возможно, другие) матрицы* $\hat{\Phi}, \hat{\Theta}$*, причём* $|\hat{T}| < a \cdot |T|$ *для некоторого* a*:*

$$n_{dw} = n_d \sum_{t \in \hat{T}} \hat{\phi}_{wt} \hat{\theta}_{td}.$$

- *для каждой пары тем* t_i, t_j *есть документ* $d_{i,j}$*, в котором используются ровно эти две темы:*

$$p(t|d_{i,j}) \neq 0 \Leftrightarrow (t = t_i \text{ или } t = t_j).$$

Тогда найдены почти все исходные темы (и, возможно, некоторые лишние):

$$|T \setminus \hat{T}| \leq a - 1.$$

Доказательство. Будем рассматривать только документы $d_{i,j}$, причём по одному для каждой пары i, j (таких документов $\binom{|T|}{2}$), и покажем, что для их полного восстановления с помощью матриц $\hat{\Phi}, \hat{\Theta}$ необходимо выполнение утверждения теоремы.

Пусть нашлась тема $\hat{t} \in \hat{T} \setminus T$, то есть являющаяся комбинацией некоторых тем из T:

$$\hat{t} = \alpha t_i + \beta t_j + \ldots, \text{ где } \alpha, \beta \neq 0.$$

Тогда $\hat{p}(\hat{t}, d) \neq 0$ не более чем для одного документа ($d = d_{i,j}$), т.к. в остальные не входит либо t_i либо t_j. Исключим из рассмотрения эту тему $\hat{t}$ и этот документ $d_{i,j}$. Повторим описанные действия k раз пока $\hat{T} \setminus T \neq \emptyset$.

Таким образом останется $|\hat{T}| - k$ тем, входящих также в T, и не менее $\binom{|T|}{2} - k$ документов. Из этих оставшихся тем можно построить только не более $\binom{|\hat{T}|-k}{2}$ документов вида $d_{i,j}$. Противоречия не будет, только если

$$\binom{|\hat{T}| - k}{2} \geq \binom{|T|}{2} - k.$$

Проведя вычисления можно получить, что

$$|\hat{T} \setminus T| = k \leq |\hat{T}| - |T| + (a - 1),$$

то есть

$$|T \setminus \hat{T}| \leq a + 1,$$

что и является утверждением теоремы. ■

Теорема 2. *Обозначим $D(T') = \{d : p(t|d) > 0 \forall t \in T'\}$ - множество документов, включающих все темы из некоторого множества T'.*

Пусть:

- *Все документы имеют одинаковую длину:*

$$n_d = \frac{n}{|D|}, \qquad p(d) = \frac{1}{|D|}.$$

- *Документы распределены равномерно внутри многогранника, ограниченного исходными темами.*

- *$c|D| \leq |D(t_i)| \leq C|D|$ и $c^{|T'|}|D| \leq |D(T')| \leq C^{|T'|}|D|$, то есть ненулевые элементы в матрице Θ распределены по столбцам независимо.*

- *В восстановленных темах есть все $|T|$ исходных тем и не более $|T|$ выпуклых комбинаций из двух тем:*

$$\hat{\Phi} = A\Phi,$$

- *причём все элементы матрицы A больше некоторого α.*

Тогда можно оценить количество M таких исходных тем t_i, что $\hat{p}(t_i) \leq \hat{p}(t')$, как

$$M \leq \frac{|T|C^2}{\alpha c - C^2}.$$

Доказательство. Пусть одна из восстановленных тем — линейная комбинация двух исходных:

$$t' = at_i + (1-a)t_j.$$

Она может входить только в документы из $D(t_i t_j)$. Рассмотрим некоторый такой документ d:

$$d = B \cdot (\beta t_i + (1-\beta)t_j) + (1-B) \cdot \dots.$$

С другой стороны, этот документ восстановлен из полученных тем:

$$d = B \cdot (\gamma' t' + \gamma_i t_i + \gamma_j t_j) + (1-B) \cdot \dots$$

Максимальное значение γ' будет в том случае, когда γ_i или γ_j нулевое. Пусть $\gamma_i = 0$, тогда приравняем и получим $\gamma' = \frac{\beta}{a}$. Итак,

$$\hat{p}(t'|d) = B\gamma' \leq B\frac{\beta}{\alpha} = \frac{p(t_i|d)}{\alpha},$$

и значит

$$\hat{p}(t') \leq \frac{1}{\alpha} \sum_{d \in D(t_i t_j)} p(t_i|d)p(d) = \frac{1}{\alpha}\frac{|D(t_i t_j)|}{|D|}\frac{|D|}{|D(t_i)|}\langle p(t_i|d)\rangle \leq \frac{C^2}{\alpha c|T|}.$$

Так как $\sum_t \hat{p}(t) = 1$, а суммарная $\hat{p}$ для всех тем-комбинаций

$$\sum_{t'} \hat{p}(t') \leq \frac{C^2}{\alpha c},$$

то суммарная $\hat{p}$ для исходных тем:

$$\sum_{t_i} \hat{p}(t_i) = 1 - \sum_{t'} \hat{p}(t') \geq 1 - \frac{C^2}{\alpha c}.$$

Оценим максимальное количество M таких t_i, что $\hat{p}(t_i) \leq \hat{p}(t') \leq \frac{C^2}{\alpha c|T|}$. Для этого заметим, что максимальное возможное значение остальных $\hat{p}(t_i)$ это $p(t_i) = \frac{1}{|T|}$. Тогда

$$1 - \frac{C^2}{\alpha c} \leq \sum_{t_i} \hat{p}(t_i) \leq (|T| - M)\frac{1}{|T|} + M\frac{C^2}{\alpha c|T|},$$

и отсюда

$$\frac{M}{|T|} \leq \frac{C^2}{\alpha c - C^2}.$$

■

В частности, если взять достаточно большое число тем $|T| = 2000$ и реалистичные значения остальных параметров $c = \frac{5}{|T|}$, $C = \frac{10}{|T|}$, $\alpha = 0.1$, то $\frac{M}{|T|} \lesssim 0.1$ — то есть, как минимум 90% исходных тем будут иметь $\hat{p}(t)$ больше, чем все возникшие линейные комбинации.

Итак, в случае выполнения предположений обеих теорем, любой метод (в частности, ARTM без регуляризаторов) восстанавливает почти все исходные темы, и в полученной модели значение $\hat{n}_t = \hat{p}(t)n$ у них выше, чем у линейных комбинаций нескольких исходных тем. В связи с этим (см. комментарий к (9)) добавление рассматриваемого регуляризатора удалит именно лишние темы с меньшими $\hat{n}_t$, оставив в итоге почти все исходные. Это улучшает интерпретируемость модели, так как смеси нескольких имеющихся тем не несут новой информации, а зашумляют результаты. Такое поведение будет иметь место для любого коэффициента регуляризации из диапазона, соответствующего различию между $\hat{n}_t$ у исходных и лишних тем — то есть, при таких τ число тем и сами темы определятся почти точно.

В ходе вычислительных экспериментов будет проверено, действительно ли такое поведение наблюдается в реальных ситуациях.

7 Вычислительный эксперимент

В качестве основы для всех экспериментов используется англоязычная коллекция статей с конференции NIPS за 12 лет. Она содержит $|D| = 1740$ документов общим размером $n \approx 2.3 \cdot 10^6$ слов, словарь содержит $|W| \approx 1.3 \cdot 10^4$ терминов. Далее эту коллекцию будем обозначать как n^1_{dw}.

Для оценки того, насколько хорошо определяется число тем, требуются коллекции с известным истинным значением. Семейство таких коллекций было сгенерировано искусственно, но на основе описанной выше коллекции статей. Было построено несколько простых моделей PLSA (без регуляризаторов) с различными фиксированными числами тем ($T_0 = 25$ и $T_0 = 50$), и синтетические коллекции n^0_{dw} получены из них следующим образом:

$$n^0_{dw} = n_d \sum_t \phi_{wt}\theta_{td},$$

где Φ и Θ — матрицы построенных моделей.

Чтобы проанализировать поведение предлагаемого метода в промежуточных случаях, рассматривается параметрическое семейство коллекций, представляющих собой смесь реальных и синтетических данных:

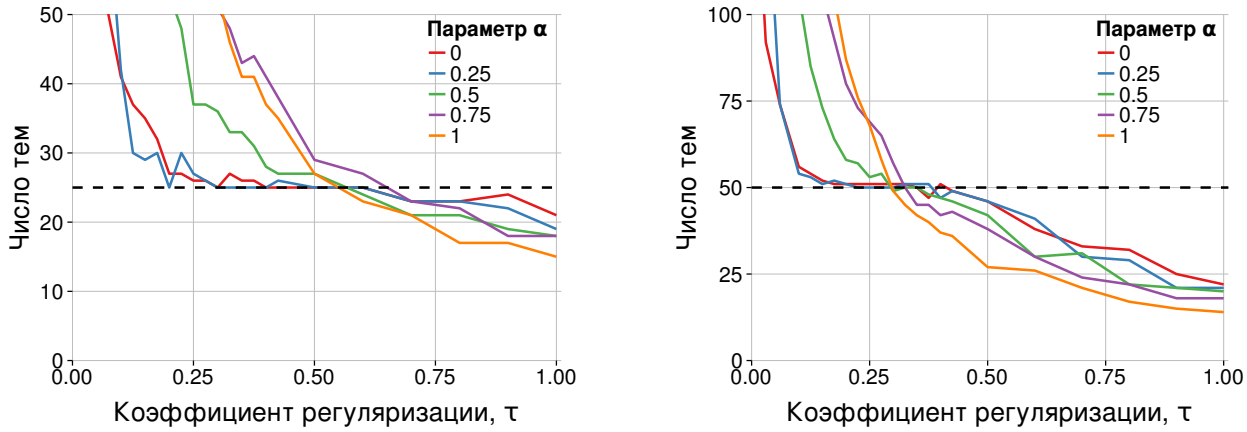
$$n^\alpha_{dw} = \alpha n^1_{dw} + (1 - \alpha)n^0_{dw}, \text{ где } \alpha \in [0, 1].$$

Следующие подразделы содержат описания экспериментов, направленных на изучение поведения предлагаемого метода на этих данных, а также их результаты.

7.1 Определение числа тем

Для определения числа тем предложенный алгоритм был запущен с различными значениями коэффициента регуляризации $\tau \in [0, 1]$. По полученным результатам построена соответствующая зависимость: рис. 2.

Здесь видно, что при полностью синтетических ($\alpha = 0$) данных, как и для близких к ним ($\alpha = 0.25$) предложенный метод верно находит истинное число тем в коллекции для широкого диапазона значений τ. В качестве рекомендуемого значе-



(a) Истинное число тем $T_0 = 25$

(b) Истинное число тем $T_0 = 50$

Рис. 2: Зависимость найденного числа тем от коэффициента регуляризации. Горизонтальной линией отмечено истинное значение.

ния параметра, по аналогии с параметром η у HDP, можно предложить $\tau = 0.3$ из этого диапазона.[1]

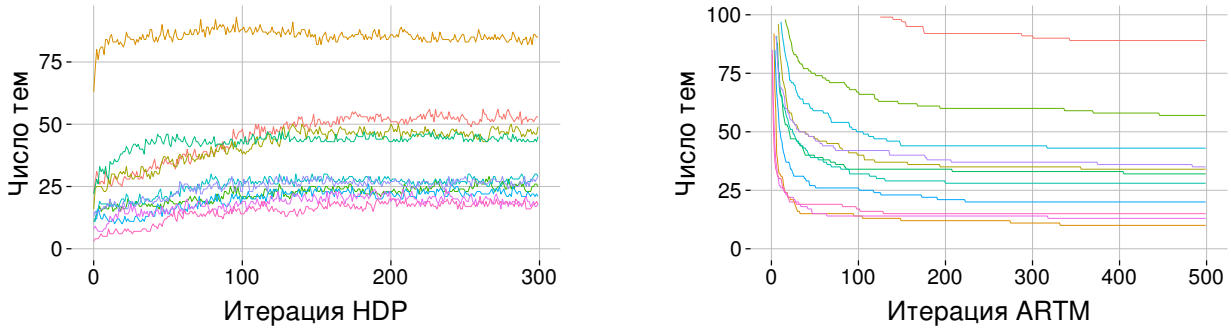
При приближении к реальным данным ($\alpha = 1$) горизонтальный участок не наблюдается, что может говорить об отсутствии истинного числа тем в таком случае. Это соответствует интуитивному представлению о том, что вообще реальную коллекцию можно успешно описывать с использованием как большого, так и маленького числа тем.

7.2 Сравнение с HDP

Для сравнения с наиболее часто используемым в аналогичных задачах алгоритмом HDP были проведены следующие эксперименты, в которых использовалась готовая реализация алгоритма HDP от C.Wang и D.Blei на языке C[2]. Он был запущен с 240 различными значениями параметра η из диапазона $[0.2, 1.7]$, который включает в себя рекомендуемое авторами значение $\eta = 0.5$. На рис. 3(a) показано изменение числа тем по ходу итераций для нескольких различных запусков HDP, и как видно, к 300 итерации результат каждый раз получается достаточно стабильным, хотя и со

[1] Что лучше сделать с этой фразой после того, как переместил этот раздел выше HDP?

[2] `http://www.cs.cmu.edu/~chongw/software/hdp.tar.gz`



(a) HDP

(b) ARTM

Рис. 3: Зависимость числа тем от итераций для нескольких запусков при различных параметрах алгоритмов

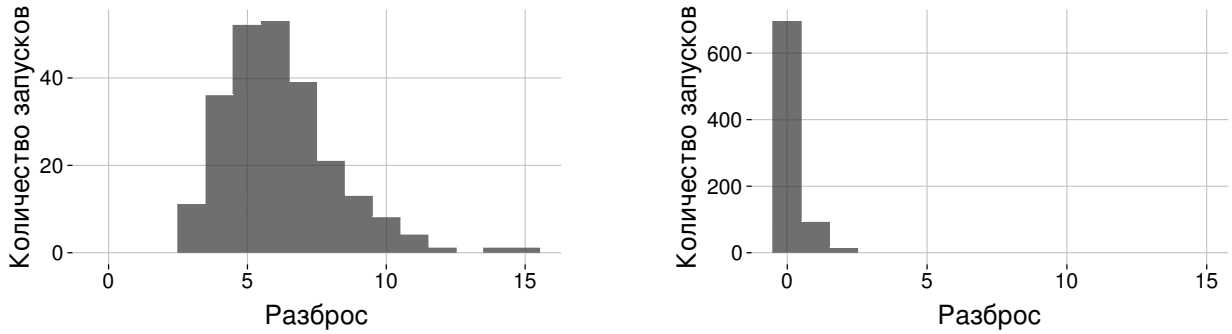
значительными колебаниями. Следовательно, такого числа итераций достаточно, и далее используется именно оно.

Библиотека, содержащая алгоритм ARTM, была реализована автором на языке Python[3]. ARTM с предлагаемым регуляризатором, который для краткости будем далее называть просто ARTM, также был запущен для большого числа значений коэффициента регуляризации, а также для нескольких начальных приближений T_{init}: 200 различных $\tau \in [0, 1.2]$ для каждого $T_{init} \in \{100, 200, 300, 500\}$. Аналогичные графики для нескольких запусков приведены на рис. 3(b), и здесь к концу процесса результат также становится достаточно стабильным.

Для такого рода алгоритмов имеет смысл рассматривать два различных вида случайного разброса получаемых значений: флуктуации внутри отдельно взятого итерационного процесса, и разброс между разными запусками с одним значением параметра. Также представляет интерес зависимость определяемого числа тем от параметра, то есть насколько будет меняться результат при отклонении от рекомендуемого авторами значения. В исследованиях HDP, как и в его применении, этот вопрос обычно не рассматривается вообще и используется $\eta = 0.5$.

Разбросом результата внутри одного запуска будем считать разность между минимальным и максимальным числом тем на последних 50 итерациях алгоритма, и отобразим эти значения в виде гистограмм для обоих рассматриваемых методов на

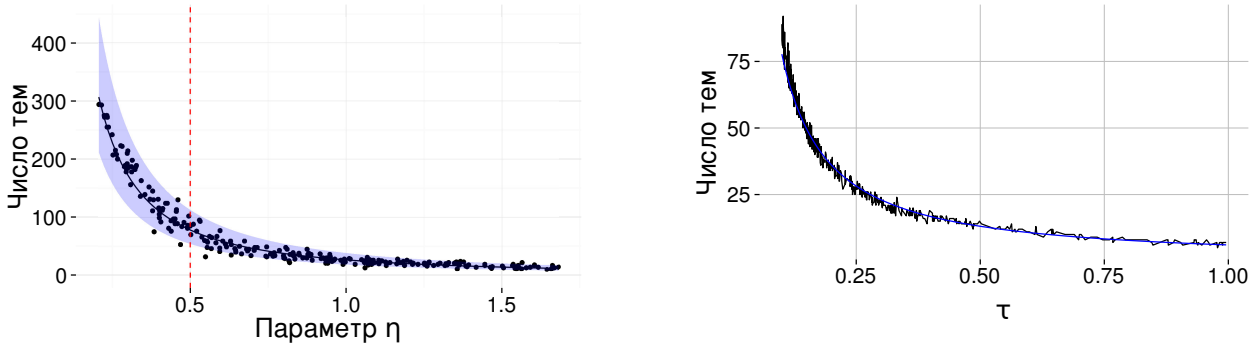
[3] `https://pypi.python.org/pypi/py-artm`



(a) HDP

(b) ARTM

Рис. 4: Разброс внутри отдельных запусков (рассматриваются последние 50 итераций)



(a) HDP. Вертикальной линией обозначено рекомендованное авторами значение параметра $\eta = 0.5$

(b) ARTM. Нормированное значение τ

Рис. 5: Зависимость найденного числа тем от параметра метода

рис. 4. Как видно, для HDP такой разброс обычно составляет 5-10 тем, в то время как в случае предлагаемого метода число тем на рассматриваемом участке процесса почти не меняется более, чем на 1 тему.

Что касается зависимости от параметра, то на рис. 5(a) показаны все полученные в эксперименте с HDP конечные результаты. Для удобства там же проведена регрессия (вида $c_1\eta^{c_2}$) с 95% доверительным интервалом. Здесь можно видеть, что существует явная обратная зависимость между параметром и определяемым числом тем, и меняя η можно получить почти любое итоговое значение. Таким образом, один параметр — число тем в LDA — заменяется на другой — η в HDP. Тем не менее, такая замена полезна: η можно задавать одинаковым для всех коллекций, в отличие от числа тем, различного в каждом случае.

В случае с ARTM помимо зависимости получаемого числа от параметра — коэффициента регуляризации τ — также существует зависимость от начального приближения для количества тем: рис. 6(a). Однако, если же нормировать $\tau_N = \tau \frac{100}{T_{init}}$ на начальное число тем T_{init} (рис. 6(b)), то результат при одном τ будет постоянным.

Объяснение такого эффекта можно получить, преобразовав формулу M-шага EM-алгоритма, на которой используется регуляризатор (9):

$$\theta_{td} = \text{norm}\left[n_{td}\left(1 - \tau\frac{1}{T_{init}p(t)}\right)_+\right] = \text{norm}\left[n_{td}\left(1 - \frac{\tau_N}{100}\frac{1}{p(t)}\right)_+\right].$$

Так как при слишком больших значениях $\tau > 1.2$ обнуление всех тем происходит на самых первых итерациях алгоритма, в этом случае $p(t) \approx \frac{1}{T_{init}}$ для всех t. Тогда

$$\theta_{td} = \text{norm}\left(n_{td}\left(1 - \tau\right)_+\right),$$

то есть результат при одинаковом τ не зависит от выбора начального значения T_{init}. Это соответствует рис. 6(a), где характерные точки графиков (в частности, обнуление числа тем при $\tau > 1.2$) не зависят от T_{init}.

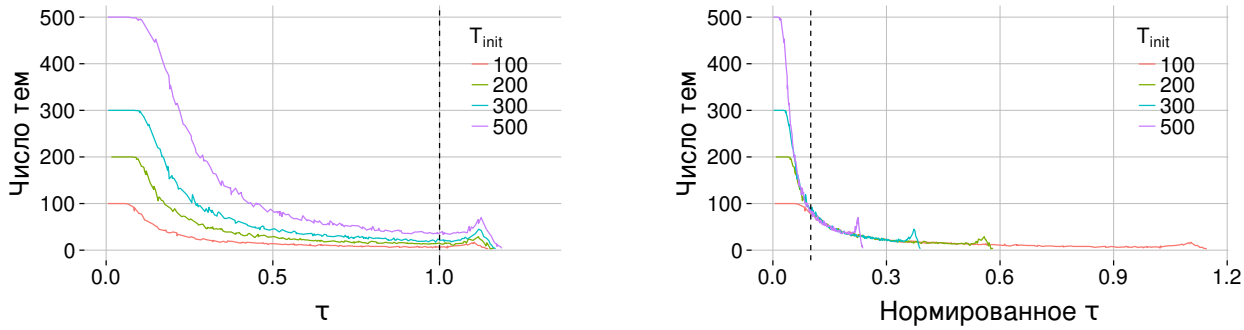
С другой стороны, если некоторым образом при различных T_{init} было достигнуто одинаковое число тем (возможно, на различных итерациях), тогда $p(t)$ в этих случаях также будут близки. Следовательно,

$$\theta_{td} = \text{norm}\left[n_{td}\left(1 - \frac{\tau_N}{100}\frac{1}{p(t)}\right)_+\right]$$

не зависит от T_{init} и дальше итерационные процессы будут идти одинаково, сходясь к одному результату, что и наблюдается на рис. 6(b).

Для дальнейшего изучения будем использовать нормированные значения τ_N и оставим только центральный участок графиков, избегая артефактов при больших и малых значениях τ (обозначены пунктирными линиями): удаление всех тем при $\tau > 1.2$ и небольшой пик перед этим значением на рис. 6(a), а также горизонтальный участок, где отбора тем не происходит на рис. 6(b).

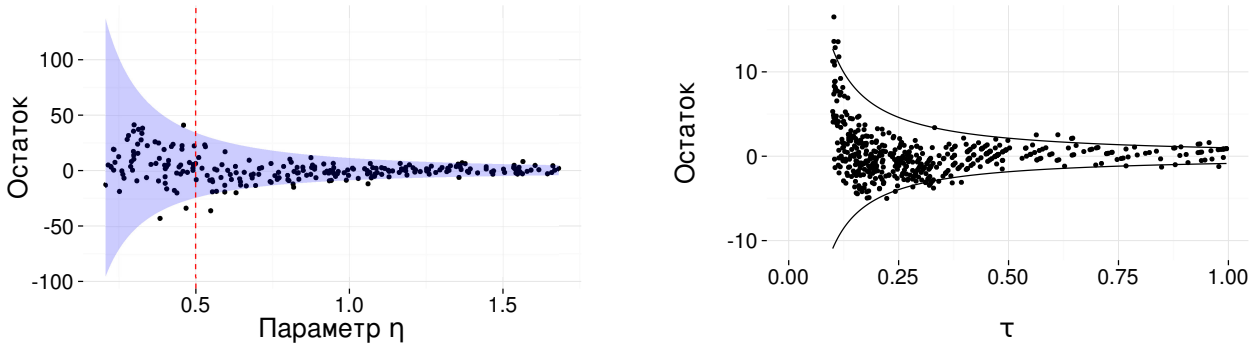
Итак, теперь аналогичным образом нанесём на график все полученные конечные результаты ARTM: рис. 5(b). Здесь также показана линия регрессии вида $c_1\eta^{c_2}$, но без доверительного интервала: как будет видно далее, здесь разброс значительно меньше, чем у HDP, и интервал было бы не видно. Общая же ситуация здесь такая же: есть явная зависимость числа тем от параметра, то есть происходит замена одного параметра на другой.



(a) Исходные значения τ

(b) Нормированные значения $\tau_N = \tau \frac{100}{T_{init}}$

Рис. 6: Зависимость числа тем от параметра в ARTM



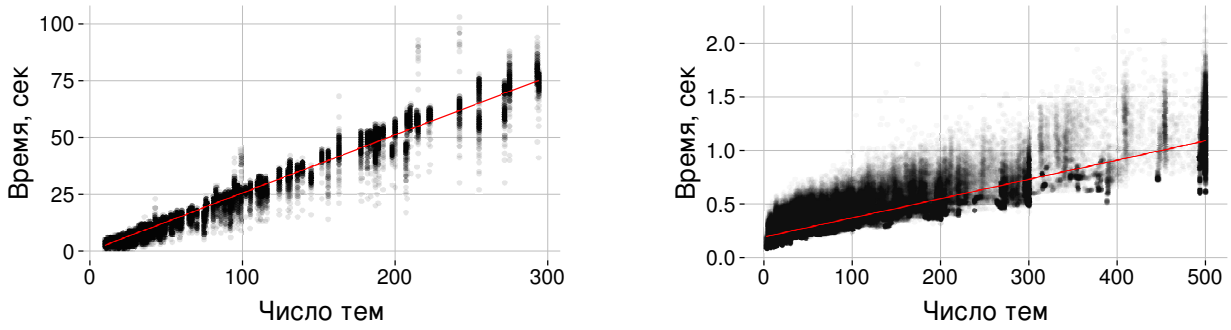
(a) HDP

(b) ARTM

Рис. 7: Разброс между разными запусками

Для оценки второго вида разброса (между отдельными запусками) удобно использовать графики на рис. 7, которые получены путём вычитания указанной регрессии из значений на рис. 5. Эти графики ясно показывают, что разброс между запусками для обоих методов значительно больше, чем внутренний. Для типичных значений параметров он составляет около 60 тем для HDP, и 5-10 тем для ARTM. Таким образом, ARTM показывает значительно более устойчивые результаты, чем ARTM.

Также имеет смысл сравнить время выполнения используемых реализаций обоих методов. Оба алгоритма были запущены в одинаковых условиях, и на рис. 8 приведены зависимости времени выполнения одной итерации от числа тем для каждого из них. Оба алгоритма используют близкое число итераций, поэтому время суммарное время работы тоже будет подчиняться подобной зависимости. Как и ожидалось, время в каждом случае пропорционально числу тем: с коэффициентом 0.25 секунд на



(a) HDP (b) ARTM

Рис. 8: Время работы алгоритмов, в расчёте на одну итерацию

	HDP	ARTM
Разброс внутри одного запуска	5-10 тем	1 тема
Разброс между запусками	60 тем	5-10 тем
Время работы (200 тем, 500 итераций)	7 часов	4.5 минуты
Возможность учёта дополнительных требований к модели	-	+

Таблица 1: Итоги сравнения HDP и ARTM

тему для HDP и 0.0018 для ARTM. Видно, что ARTM работает значительно быстрее, в частности при 200 темах и 500 итерациях прирост скорости составляет около 100 раз: 7 часов для HDP против 4.5 минут для ARTM.

Наконец, приведём результаты сравнения этих двух методов в таблице 1.

7.3 Удаление смесей и долей тем

Для получения полезных интерпретируемых тем важно, чтобы в полученной модели не было тем, являющихся смесями или долями уже имеющихся. Чтобы проверить, насколько хорошо предлагаемый метод удаляет именно лишние темы, в рассмотренную ранее синтетическую коллекцию n_{dw}^0 с 50 темами было искусственно добавлено T_1 дополнительных тем. Эти темы были сгенерированы двумя разными способами:

- Как выпуклые комбинации имеющихся тем:

$$\phi_{wt} = \sum_{t' \leq T_0} \alpha_{tt'} \phi_{wt'}, \; t \in [T_0 + 1, T_0 + T_1],$$

где α — матрица из $T_1 \times T_0$ случайных коэффициентов с T_{comb} ненулевыми элементами в каждой строке, каждая строка которой нормирована: $\sum_{t'} \alpha_{tt'} = 1$.

Распределение этих тем по документам было произведено случайным заполнением соответствующих строк матрицы Θ. Случайные значения брались из уже существующих строк для того, чтобы распределения $p(t|d)$ для исходных и новых тем были аналогичны.

- Наоборот, как доли имеющихся тем:

$$\phi_{wt} = \sum_{T_0 < t' \leq T_0 + T_1} \alpha_{tt'} \phi_{wt'},\ t \in [1, T_0],$$

где α — матрица из $T_0 \times T_1$ случайных коэффициентов с T_{comb} ненулевыми элементами в каждой строке, каждая строка которой нормирована: $\sum_{t'} \alpha_{tt'} = 1$.

В этом случае распределения $p(t'|d)$ выбирались такими же, как $p(t|d)$ у исходной темы с добавлением шума:

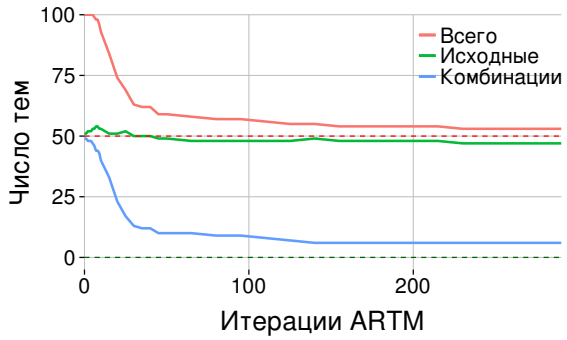
$$p(t'|d) = (1 + \varepsilon) p(t|d).$$

Результаты почти не зависят от размера шума, поэтому в описанных экспериментах используется $\varepsilon \in [-0.3, 0.3]$.

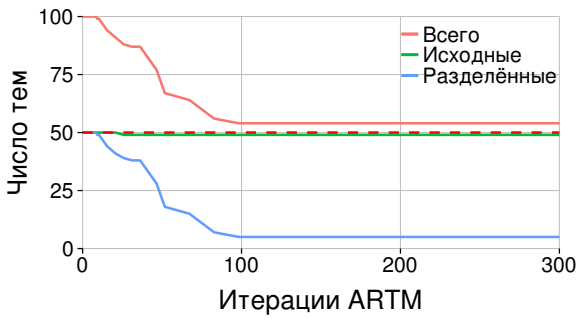
Здесь в первом случае лишними оказываются добавленные комбинации тем, так как они полностью могут быть заменены исходными темами, но не наоборот. Во втором случае распределения $p(t|d)$ у разделённых тем такие же, как у соответствующих исходных, и они не требуются для описания коллекции, поэтому ситуация противоположная: лишними являются добавленные разделённые темы.

На таких коллекциях был запущен ARTM с предлагаемым регуляризатором, причём коэффициент регуляризации использовался близкий к оптимальному ($\tau = 0.3$, см. предыдущий раздел). На каждой итерации записано, сколько осталось исходных тем, и сколько добавленных: по одному примеру таких запусков с $T_1 = 50$ и $T_{comb} = 5$ приведено на рис. 9. Как видно, к концу остаются в основном исходные темы и почти все лишние удаляются.

Такие эксперименты были проведены для $T_1 \in \{10, 20, 50, 100\}$ и для $T_{comb} \in \{2, 5, 20\}$, и на рис. 10 показана эффективность метода с рассматриваемой точки

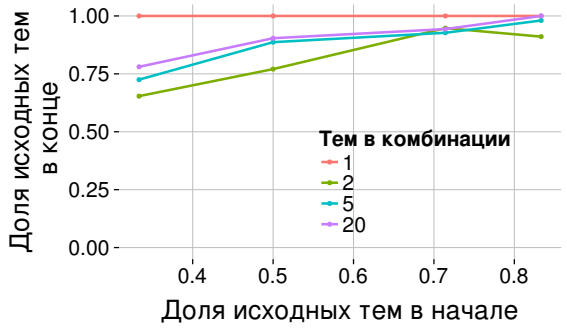


(a) Добавлено 50 комбинаций по 5 тем

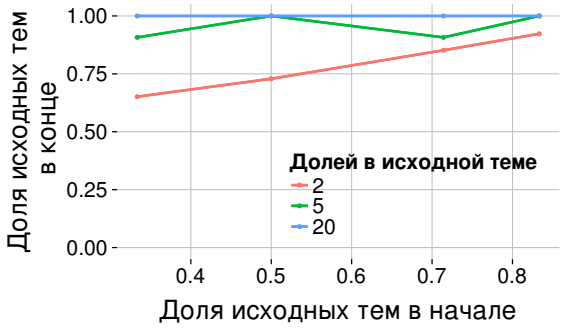


(b) Добавлено 50 тем, полученных разделением 10 исходных по 5 долей

Рис. 9: Примеры запусков с удалением лишних тем



(a) Добавлены комбинации тем



(b) Добавлены разделённые темы

Рис. 10: Удаление лишних тем

зрения в этих случаях. Как видно, во всех случаях преимущественно удаляются именно лишние темы, хотя и несколько хуже в случае большого числа лишних тем и меньших T_{comb}.

8 Заключение

В работе предложен метод отбора тем в модели ARTM на основе энтропийного регуляризатора. Показано, что он позволяет определять число тем в коллекции значительно устойчивее и быстрее, чем стандартный метод HDP. Показано, что в первую очередь он удаляет лишние для модели комбинации тем и расщеплённые темы. Доказано и показано экспериментально, что в случае, когда истинное число тем в коллекции существует, предлагаемый метод позволяет определять его.

В дальнейших исследованиях возможна оптимизация коэффициента регуляризации на основе некоторой метрики качества, а также разработка более общего метода, сочетающего поочерёдный отбор тем и добавление новых.

Список литературы

[1] *Yi X., Allan J.* A comparative study of utilizing topic models for information retrieval // Advances in Information Retrieval. — 2009. — Pp. 29–41.

[2] Statistical topic models for multi-label document classification / T. N. Rubin, A. Chambers, P. Smyth, M. Steyvers // *Mach. Learn.* — 2012. — . — Vol. 88, no. 1-2. — Pp. 157–208. http://dx.doi.org/10.1007/s10994-011-5272-5.

[3] *Yeh J.-h., Wu M.-l.* Recommendation based on latent topics and social network analysis // Computer Engineering and Applications (ICCEA), 2010 Second International Conference on / IEEE. — Vol. 1. — 2010. — Pp. 209–213.

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

[5] *Hofmann T.* Probabilistic latent semantic indexing // Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. — SIGIR '99. — New York, NY, USA: ACM, 1999. — Pp. 50–57. http://doi.acm.org/10.1145/312624.312649.

[6] *Воронцов К. В.* Аддитивная регуляризация тематических моделей коллекций текстовых документов // Доклады РАН. — Vol. 455. — 2014.

[7] *Vorontsov K., Potapenko A., Plavin A.* Additive regularization of topic models for topic selection and sparse factorization // SLDS 2015. — 2015.

[8] *Blei D. M., Ng A. Y., Jordan M. I.* Latent dirichlet allocation // *J. Mach. Learn. Res.* — 2003. — . — Vol. 3. — Pp. 993–1022. http://dl.acm.org/citation.cfm?id=944919.944937.

[9] Hierarchical dirichlet processes / Y. W. Teh, M. I. Jordan, M. J. Beal, D. M. Blei // *Journal of The American Statistical Association.* — 2006.

[10] *Tang J., Zhang M., Mei Q.* "look ma, no hands!"A parameter-free topic model // *CoRR.* — 2014. — Vol. abs/1409.2993. http://arxiv.org/abs/1409.2993.

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

[12] *Vorontsov K. V., Potapenko A. A.* Tutorial on probabilistic topic modeling: Additive regularization for stochastic matrix factorization // Analysis of Images, Social networks and Texts. — 2014.

[13] *Gupta M. R., Chen Y.* Theory and Use of the EM Algorithm. — Now Publishers Inc, 2011.

[14] *Воронцов К. В., Потапенко А. А.* Модификации EM-алгоритма для вероятностного тематического моделирования // *Машинное обучение и анализ данных.* — 2013.

[15] *Воронцов К. В.* Вероятностное тематическое моделирование. — 2014.

[16] A sparsity constraint for topic models-application to temporal activity mining / J.-M. Odobez, J. Varadarajan, R. Emonet et al. // NIPS-2010 Workshop on Practical Applications of Sparse Modeling: Open Issues and New Directions. — No. EPFL-CONF-155013. — 2010.

[17] *Kullback S., Leibler R. A.* On information and sufficiency // *The Annals of Mathematical Statistics.* — 1951. — Pp. 79–86.