Министерство образования и науки Российской Федерации
Московский физико-технический институт (государственный университет)
Факультет управления и прикладной математики
Вычислительный центр им. А. А. Дородницына РАН
Кафедра «Интеллектуальные системы»

Захаренков Антон Александрович

Итерационный подбор коэффициентов регуляризации в тематическом моделировании

03.03.01 — Прикладные математика и физика

Выпускная квалификационная работа бакалавра

Научный руководитель:
Профессор РАН, д. ф.-м. н.
Воронцов Константин
Вячеславович

Москва
2018

Содержание

Часть I

Введение

Актуальность темы. Тематическое моделирование - одно из наиболее актуальных и быстро развивающийхся направлений в анализе естественного языка. Вероятностная тематическая модель выявляет тематику коллекции текстовых документов, описывая каждую тему дискретным распределением на множестве терминов, каждый документ — дискретным распределением на множестве тем. Таким образом, тематическая модель выступает в качестве средства систематизации и анализа информации в больших текстовых коллекциях.

Цели работы.

- применение существующих алгоритмов глобальной оптимизации к модели bigARTM;
- модификация существующих алгоритмов глобальной оптимизации с учетом специфики модели bigARTM;
- сравнение работы этих алгоритмов на реальных данных.

1 Вероятностное тематическое моделирование

Тематическое моделирование — это одно из современных направлений статистического анализа текстов, активно развивающееся с конца 90-х годов. Приложения тематических моделей — информационный поиск, нахождение основных трендов в научных публикациях или новостных потоках [3, 4], для классификации и категоризации [5, 6] документов, изображений и видеопотоков [7, 8, 9, 10], для тематической сегментации текстов [11], для информационного поиска [12, 13, 14, 15, 16, 17, 18, 19], в том числе многоязычного [20, 21], для тегирования веб-страниц [22], для анализа данных социальных сетей [23, 24, 25], для обнаружения текстового спама [26], для рекомендательных систем [27, 11, 28, 29, 30], для анализа нуклеотидных [31] и аминокислотных последовательностей [32, 33], в задачах популяционной генетики [34] и других приложений из разных областей.

1.1 Вероятностная постановка задачи

Задано множество слов W (словарь). Задано множество текстовых документов D (коллекция), каждый документ которого $d \in D$ является упорядоченным подмультимножеством из W и состоит из n_d слов w, которые могут повторяться.

Необходимо сделать несколько важных предположений:

1. предполагается, что существует конечное множество скрытых тем T, и каждое употребление термина w в каждом документе d связано с некоторой темой $t \in T$, которая не известна;

2. предполагается, что коллекция документов — это множество троек (d, w, t), выбранных случайно и независимо из дискретного распределения $\mathsf{P}(d, w, t)$, заданного на конечном множестве $D \times W \times T$ (T — фиксировано);

3. предполагается, что порядок терминов в документах не важен для выявления тематики. Эту гипотезу ещё называют «мешком слов» (bag of words);

4. предполагается, что порядок документов в теме также не важен. Эту гипотезу называют «мешком документов» (bag of documents).

В данных предположениях можно представлять документы как подмножества $d \subset W$, в которых каждому элементу $w \in d$ поставлено в соответствие число его вхождений в документ n_{dw}, т.е. значения признаков $f_w(d) = n_{dw}$ для всех слов $w \in W$ документа $d \in D$.

Таким образом, построить тематическую модель коллекции документов D — значит найти множество тем T, распределения слов $\mathsf{P}(w\,|\,t)$ для всех тем $t \in T$ и распределения тем $\mathsf{P}(t\,|\,d)$ для всех документов $d \in D$. В данной постановке термин может принадлежать к нескольким темам.

1.2 Вероятностная модель коллекции документов

Для вероятностной модели вводится ещё несколько предположений. *Гипотеза условной независимости* предполагает, что независимость появления слов в документе d, относящихся к теме t, описывается общим для всей коллекции распределением $\mathsf{P}(w\,|\,t)$ и не зависит от документа d, т.е. $\mathsf{P}(w\,|\,t,d) = \mathsf{P}(w\,|\,t)$.

Тогда по формуле полной вероятности получаем модель порождения данных по распределениям тем в документах и слов в темах:

$$\mathsf{P}(w\,|\,d) = \sum_{t \in T} \mathsf{P}(t\,|\,d)\,\mathsf{P}(w\,|\,t). \tag{1.1}$$

Тематическое моделирование рассматривает обратную задачу: по известной коллекции документов D требуется восстановить породившие её распределения.

Также предполагается, что количество тем n_t не очень большое, и

задачу (1.1) можно представить, как задачу разложения матрицы частот

$$F_{WD} = (\mathsf{P}(w\,|\,d))_{W\times D}, \quad \mathsf{P}(w\,|\,d) = \frac{n_{dw}}{n_d},$$

на произведение двух неизвестных матриц:

$$F_{WD} \;\approx\; \Phi_{WT} \;\times\; \Theta_{TD}; \tag{1.1′}$$
$$\Phi_{WT} = (\varphi_{wt})_{W\times T}, \quad \varphi_{wt} = \mathsf{P}(w\,|\,t) = \frac{n_{wt}}{n_t};$$
$$\Theta_{TD} = (\theta_{td})_{T\times D}, \quad \theta_{td} = \mathsf{P}(t\,|\,d) = \frac{n_{td}}{n_d}.$$

Причём существенно, что матрицы F_{WD}, Φ_{WT}, Θ_{TD} — *стохастические*, т.е. имеют неотрицательные нормированные столбцы, поэтому применить произвольный метод представления матрицы в виде произведения двух неизвестных матриц не получится.

1.3 Вероятностный латентный семантический анализ

В *вероятностном латентном семантическом анализе* (PLSA) [39] для построения модели (1.1′) предлагается максимизировать логарифм правдоподобия (плотности распределения) выборки при ограничениях неотрицательности и нормировки столбцов матриц Φ и Θ:

$$L(\Phi,\Theta) = \ln\prod_{d\in D}\prod_{w\in d}\mathsf{P}(w\,|\,d)^{n_{dw}} = \sum_{d\in D}\sum_{w\in d} n_{dw}\ln\sum_{t\in T}\varphi_{wt}\theta_{td} \to \max_{\Phi,\Theta}; \tag{1.2}$$
$$\sum_{w\in W}\varphi_{wt} = 1, \quad \varphi_{wt}\geqslant 0; \qquad \sum_{t\in T}\theta_{td} = 1, \quad \theta_{td}\geqslant 0.$$

Максимум задачи (1.2) в вероятностном латентном семантическом анализе ищут с помощью итерационного алгоритма EM [40]. Этот алгоритм заключается в чередовании двух шагов E-шага (expectation) и M-шага (maximization).

На E-шаге алгоритм по текущим значениям $\varphi_{wt}, \theta_{td}$ вычисляет услов-

ные вероятности всех тем для каждой пары термин-документ:

$$H_{dwt} = \mathsf{P}(t \,|\, d, w) = \frac{\mathsf{P}(w \,|\, t)\,\mathsf{P}(t \,|\, d)}{\mathsf{P}(w \,|\, d)}. \tag{1.3}$$

На М-шаге, если принять оценку n_{dwt}:

$$n_{dwt} \approx n_{dw}\,\mathsf{P}(t \,|\, d, w) = n_{dw} H_{dwt}, \tag{1.4}$$

несложно по условным вероятностям пересчитать новое приближение параметров:

$$\varphi_{wt} = \frac{n_{wt}}{n_t}, \qquad n_t = \sum_{w \in W} n_{wt}, \qquad n_{wt} = \sum_{d \in D} n_{dwt}; \tag{1.5}$$

$$\theta_{td} = \frac{n_{dt}}{n_d}, \qquad n_d = \sum_{t \in T} n_{dt}, \qquad n_{dt} = \sum_{w \in d} n_{dwt}. \tag{1.6}$$

ЕМ-алгоритм находит стационарную точку функционала при заданных положительных условиях.

Лагранжиан задачи (1.2) при ограничениях нормировки выглядит следующим образом:

$$\mathscr{L} = \sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \varphi_{wt} \theta_{td} - \sum_{t \in T} \lambda_t \left(\sum_{w \in W} \varphi_{wt} - 1 \right) - \sum_{d \in D} \mu_d \left(\sum_{t \in T} \theta_{td} - 1 \right).$$

Если его продифференцировать и приравнять к нулю, то получим:

$$\frac{\partial \mathscr{L}}{\partial \varphi_{wt}} = \sum_{d \in D} n_{dw} \frac{\theta_{td}}{\mathsf{P}(w \,|\, d)} - \lambda_t = 0. \tag{1.7}$$

Умножая обе части на φ_{wt} и просуммировав обе части по всем словам, выражение принимает вид:

$$\sum_{w \in W} \varphi_{wt} \lambda_t \equiv \lambda_t = \sum_{d \in D} \sum_{w \in W} n_{dw} H_{dwt}.$$

Если снова умножить обе части (1.7) на φ_{wt} и выразить φ_{wt}, получаем:

$$\varphi_{wt} = \frac{\sum\limits_{d \in D} n_{dw} H_{dwt}}{\sum\limits_{d \in D} \sum\limits_{u \in W} n_{du} H_{dut}} \equiv \frac{n_{wt}}{n_t}, \text{ для всех } w \in W,\ t \in T.$$

Несложно заметить, что если изначальные приближения φ_{wt} и θ_{td} были положительные, то они такими и останутся после любой итерации несмотря на то, что условие неотрицательности было проигнорировано. Получаем алгоритм разложения исходной стохастической матрицы F_{WD} на произведение двух матриц меньшего размера Φ_{WT} и Θ_{TD}.

Приведём псевдокод EM-алгоритма 1 для модели PLSA, описанный в работе [1].

Algorithm 1 EM-алгоритм для модели PLSA

Вход: $D_{WD} \equiv \{d_i\}_{i=1}^{\ell}$, число тем $|T|$, начальные приближения $\Theta,\ \Phi$;
Выход: конечные матрицы распределений $\Theta,\ \Phi$;

повторять
  обнулить n_{wt}, n_{dt}, n_t для всех $w \in W,\ d \in D,\ t \in T$;
  для всех $d \in D,\ w \in d$:
    $Z := \sum\limits_{t \in T} \varphi_{wt} \theta_{td}$;
    для всех $t \in T$ таких, что $\varphi_{wt}\theta_{td} > 0$:
      $\delta := n_{dw} \varphi_{wt} \theta_{td} / Z$;
      $n_{wt} := n_{wt} + \delta$;
      $n_{dt} := n_{dt} + \delta$;
      $n_t := n_t + \delta$;
    end для
  end для
  $\varphi_{wt} := n_{wt} / n_t$ для всех $w \in W,\ t \in T$;
  $\theta_{td} := n_{dt} / n_d$ для всех $d \in D,\ t \in T$;
пока Θ, Φ не сойдутся;

Нетрудно заметить, что вычисление переменных n_{wt}, n_{dt}, n_t на M-шаге требует лишь однократного прохода всей коллекции в цикле по всем документам $d \in D$ и всем терминам $w \in d$, а переменные p_{tdw} можно вычислять только в тот момент, когда они нужны. Таким образом,

E-шаг встраивается внутрь M-шага без дополнительных вычислительных затрат, и отпадает необходимость хранения трёхмерной матрицы p_{tdw}. Этот вариант реализации EM-алгоритма принято называть рациональным. Существует множество версий EM-алгоритма, различающихся частотой обновления параметров модели φ_{wt} и θ_{td} по переменным n_{wt} и n_{td}. Частые обновления повышают скорость сходимости и слабо влияют на значение правдоподобия в конце итераций [41]. Ниже приведён псевдокод рационального EM-алгоритма 2 для модели PLSA.

Algorithm 2 рациональный EM-алгоритм для модели PLSA

Вход: $D_{WD} \equiv \{d_i\}_{i=1}^{\ell}$, число тем $|T|$, начальные приближения $\Theta,\ \Phi$;
Выход: конечные матрицы распределений $\Theta,\ \Phi$;

повторять
  обнулить n_{wt}, n_{dt}, n_t для всех $w \in W,\ d \in D,\ t \in T$;
  для всех $d \in D,\ w \in d$:
    $Z := \sum_{t \in T} \varphi_{wt}\theta_{td}$;
    для всех $t \in T$ таких, что $\varphi_{wt}\theta_{td} > 0$:
      $\delta := n_{dw}\varphi_{wt}\theta_{td}/Z$;
      $n_{wt} := n_{wt} + \delta$;
      $n_{dt} := n_{dt} + \delta$;
      $n_t := n_t + \delta$;
    end для
  end для
  $\varphi_{wt} := n_{wt}/n_t$ для всех $w \in W,\ t \in T$;
  $\theta_{td} := n_{dt}/n_d$ для всех $d \in D,\ t \in T$;
пока Θ, Φ не сойдутся;

В секции **??** будет показано, как сделать онлайн-версию EM-алгоритма, что позволит обрабатывать огромные коллекции. Именно Online EM-алгоритм [42, 43], а точнее его параллельная реализация в BigARTM [44, 45], будет использоваться при постановке экспериментов в части III.

2 Аддитивная регуляризация тематических моделей

Искомое стохастическое матричное разложение $F \approx \Phi\Theta$ в тематической модели определено не единственным образом. Задача тематического моделирования, в виду этого, имеет бесконечно много решений. Такие задачи принято называть *некорректно поставленными* [46], а общий подход к устранению данной проблемы — *регуляризацией* [47]. Её идея заключается в добавлении к логарифму правдоподобия (1.2) штрафного слагаемого, которое сужает множество решений.

Построение многоцелевых тематических моделей [48] существенно упрощается благодаря аддитивности регуляризаторов, причём добавление регуляризатора требует лишь небольшой модификации М-шага.

2.1 Постановка и решение задачи

Предположим, что вместе с логарифмом правдоподобия (1.2) требуется максимизировать ещё r критериев $R_i(\Phi,\Theta)$, $i = 1,\dots,r$, называемых регуляризаторами. Тогда для оптимизации будем рассматривать линейную комбинацию критериев $L(\Phi,\Theta)$ и $R_i(\Phi,\Theta)$ с неотрицательными *коэффициентами регуляризации* $\hat{\tau}_i$:

$$R(\Phi,\Theta) = \sum_{i=1}^{r} \hat{\tau}_i R_i(\Phi,\Theta), \quad L(\Phi,\Theta) + R(\Phi,\Theta) \to \max_{\Phi,\Theta}. \tag{2.1}$$

Решение этой задачи приводит к обобщению формул М-шага [35] в ЕМ-алгоритме:

$$\varphi_{wt} = \frac{\Big(n_{wt} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}}(\Phi,\Theta)\Big)_+}{\sum\limits_{u\in W}\Big(n_{ut} + \varphi_{ut}\dfrac{\partial R}{\partial \varphi_{ut}}(\Phi,\Theta)\Big)_+}, \quad \theta_{td} = \frac{\Big(n_{dt} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}(\Phi,\Theta)\Big)_+}{\sum\limits_{s\in T}\Big(n_{ds} + \theta_{sd}\dfrac{\partial R}{\partial \theta_{sd}}(\Phi,\Theta)\Big)_+}, \tag{2.2}$$

где $(x)_+ = \max\{0,x\}$, а значения n_{wt}, n_{dt} определяются формулами (1.3)–(1.6).

Иногда знаменатель формул опускают и заменяют нормировкой по переменной p — $\underset{p \in P}{\text{norm}}$.

Таким образом, EM-алгоритм для обучения регуляризованной модели может быть реализован путём незначительной модификации любого имеющегося EM-подобного алгоритма, а модель вероятностного латентного семантического анализа PLSA соответствует частному случаю, когда регуляризаторы отсутствуют.

В байесовских методах обучения тематических моделей [49, 50, 51] регуляризатор $R(\Phi, \Theta)$ интерпретируется как логарифм априорного распределения, а оптимизационная задача (2.1) соответствует принципу максимума апостериорной вероятности. В АРТМ регуляризатор не обязан иметь вероятностную интерпретацию.

3 Примеры регуляризаторов

Регуляризаторы R_i выписываются из соображений удобства решения регуляризованной оптимизационной задачи, чтобы в некорректно поставленных задачах достигать оптимум, который бы обладал целевыми значениями показателей качества Q_j. Обычно регуляризаторы являются гладкими функциями от матриц Φ и Θ.

Нетрудно заметить, что с точки зрения регуляризаторов можно пересмотреть разработанные в рамках байесовского подхода тематические модели и подобрать к ним подходящие (или очень близкие) регуляризаторы.

Ниже предлагаются примеры регуляризаторов, которые будут использоваться в экспериментах. Сначала приведено два важных примера общих регуляризаторов на основе дивергенции Кульбака-Лейблера, а затем примеры регуляризаторов для выделения предметных и фоновых тем и регуляризатор, увеличивающий различие тем.

О многих других регуляризаторах для АРТМ можно посмотреть в работах [52, 35, 37, 38].

3.1 Сглаживающий регуляризатор

Рассмотрим дивергенцию Кульбака-Лейблера (относительную энтропию) для двух дискретных распределений $(p_i)_{i=1}^n$ и $(q_i)_{i=1}^n$:

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

Эта функция расстояния неотрицательна, но при этом не является симметричной. Тем не менее, минимизация KL-дивергенции эквивалентна максимизации правдоподобия модели распределения q по эмпирическому распределению p.

Если задать дискретные распределения на множестве терминов $\beta = (\beta_w)_{w\in W}$ и на множестве тем $\alpha = (\alpha_t)_{t\in T}$, то можно, минимизируя суммы KL-дивергенций, добиться схожести распределений φ_t с β и θ_d с α:

$$\sum_{t\in T} KL_w(\beta\,\|\,\varphi_{wt}) \to \min_{\Phi}, \qquad \sum_{d\in D} KL_t(\alpha_t\,\|\,\theta_{td}) \to \min_{\Theta}.$$

Если переписать эти функционалы через общую максимизацию правдоподобия с коэффициентами α_0 и β_0, то получим:

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

При этом формулы М-шага приобретают вид:

$$\varphi_{wt} = \underset{w\in W}{\operatorname{norm}} \left(n_{wt} + \beta_0\beta_w\right), \qquad \theta_{td} = \underset{t\in T}{\operatorname{norm}} \left(n_{dt} + \alpha_0\alpha_t\right).$$

Данный регуляризатор принято называть *сглаживающим регуляризатором Дирихле*, т.к. он эквивалентен предположению, что столбцы матриц Φ и Θ порождаются априорными распределениями Дирихле с гиперпараметрами $\beta_0\beta_w$ и $\alpha_0\alpha_t$ [1].

3.2 Разреживающий регуляризатор

Предполагается, что каждый документ $d \in D$ и каждый термин $w \in W$ связан с небольшим числом тем $t \in T$. Это естественное предположение, т.к. если термин принадлежит большому числу тем, то он является общеупотребительным, а значит, не поможет определить тематику. Аналогично, если документ принадлежит большому числу тем, то он похож на энциклопедию и его лучше разбить на тематические части. Таким образом, большая часть вероятностей φ_{wt} и θ_{td} обнуляется, тогда как при построении моделей с большим числом тем сильная разреженность матриц помогает сократить время и память.

Заметим, что чем сильнее разрежено распределение, тем меньше его энтропия. Тогда предлагается максимизировать KL-дивергенцию между распределениями φ_t и θ_d и равномерными распределениями $\beta = (\beta_w)_{w \in W}$, $\alpha = (\alpha_t)_{t \in T}$, т.к. равномерное распределение обладает максимальной энтропией:

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

При этом формулы M-шага приобретают вид:

$$\varphi_{wt} = \underset{w \in W}{\operatorname{norm}} \left(n_{wt} - \beta_0 \beta_w \right), \qquad \theta_{td} = \underset{t \in T}{\operatorname{norm}} \left(n_{dt} - \alpha_0 \alpha_t \right).$$

Данный регуляризатор принято называть *разреживающим регуляризатором*. Он является противоположностью по оптимизационной задаче сглаживающего регуляризатора [1].

3.3 Декоррелирующий регуляризатор

Тематические модели интересны при различии тем. Формализовать понятие различности тем можно по-разному, например, как ковариацию на нормированных векторах $\varphi_w = (\varphi_{wt})_{t \in T}$ дискретных распределений $\varphi_{wt} = \mathrm{P}(w \,|\, t)$:

$$R(\Phi,\Theta) = \frac{\tau}{2}\sum_{t\in T}\sum_{s\in T\setminus t}\operatorname{cov}(\varphi_t,\varphi_s) \to \max, \quad \operatorname{cov}(\varphi_t,\varphi_s) = \sum_{w\in W}\varphi_{wt}\varphi_{ws}.$$

При этом формула φ_{wt} принимает вид:

$$\varphi_{wt} = \underset{w\in W}{\operatorname{norm}}\left(n_{wt} - \tau\varphi_{wt}\sum_{s\in T\setminus t}\varphi_{ws}\right).$$

Текущее выражение уменьшает условные вероятности $\varphi_{wt} = \mathsf{P}(w\,|\,t)$ для слов w, которые имеют большие значения вероятности φ_{ws} в других темах и увеличивают условные вероятности φ_{wt} наиболее значимых тем слова w. Таким образом, данный регуляризатор также является разреживающим. Кроме того, регуляризатор декоррелирования обладает свойством группировки слов общей лексики в отдельные темы [52, 53].

3.4 Выделение предметных и фоновых тем

Для того, чтобы тему можно было хорошо интерпретировать, она должна содержать лексическое ядро — множество слов, характерных для определённой предметной области, которые часто употребляются рядом в документах, относятся к данной теме и практически не употребляются в других темах. При этом существуют и совсем неинформативные темы, которые состоят из общих слов и не привязаны к конкретной тематике. Таким образом, множество тем разбивается на два подмножества, $T = S \sqcup G$, где S — это предметные темы, а G — фоновые темы.

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

Фоновые темы $t \in G$ содержат слова общей лексики, которых не должно быть в предметных темах. Их распределения $\mathsf{P}(w\,|\,t)$ и $\mathsf{P}(d\,|\,t)$ сглажены, так как эти слова и темы присутствуют в большинстве доку-

ментов.

4 Примеры метрик качества

Оценивание качества тематических моделей является отдельной сложной задачей. Основная сложность заключается в том, что в отличие от типичных задач с учителем, здесь нет истинных целевых меток, поэтому сложно вводить адекватные функционалы качества. А критерии качества задач без учителя, такие как среднее внутрикластерное или межкластерное расстояние, плохо подходят для оценивания совместной кластеризации документов и терминов. При этом, критерии качества могут вычисляться сложным образом, быть негладкими или многократно изменяться в ходе исследования.

Критерии качества тематических моделей принято делить на внутренние (intrinsic) и внешние (extrinsic) [52]. Внутренние критерии характеризуют качество модели по исходной текстовой коллекции и полученным результатам, а внешние критерии оценивают полезность модели путём сбора дополнительных данных с конечных пользователей.

Ниже предлагаются примеры внутренних метрик качества, которые будут использоваться в экспериментах. Сначала приведена одна из наиболее распространённых внутренних метрик качества — перплексия, которая используется во многих областях, а дальше — метрики качества, описывающие разреженность полученных матриц и характеристики ядер тем.

О многих других метриках качества для APTM можно посмотреть в работах [52, 35, 37, 38].

4.1 Перплексия

Перплексия — это мера несоответствия модели $\mathsf{P}(w\,|\,d)$ словам w. Она определяется через логарифм правдоподобия (1.2):

$$\mathscr{P}(D;p) = \exp\left(-\frac{1}{n}L(\Phi,\Theta)\right) = \exp\left(-\frac{1}{n}\sum_{d\in D}\sum_{w\in W} n_{dw}\ln \mathsf{P}(w\,|\,d)\right), \quad (4.1)$$

где $n = \sum_{d\in D}\sum_{w\in W} n_{dw}$ — длина коллекции.

Чем меньше величина перплексии, тем лучше модель p предсказывает появление токенов w в документах d коллекции D. При этом, если термины w порождаются равномерным распределением $\mathsf{P}(w) = 1/|W|$ на словаре мощности $|W|$, то перплексия модели $\mathsf{P}(w)$ на таком тексте с ростом длины словаря стремится к $|W|$. Причём, чем сильнее распределение $\mathsf{P}(w)$ отличается от равномерного, тем меньше перплексия. В случае условных вероятностей $\mathsf{P}(w\,|\,d)$, если каждый документ генерируется из $|W|$ равновероятных терминов (возможно, различных в разных документах), то перплексия сходится к $|W|$.

Недостаток данной метрики заключается в том, что конкретные численные значения перплексии не всегда очевидны. Более того, её значение также зависит от длины документов, мощности и разреженности словаря. Например, с помощью перплексии некорректно сравнивать тематические модели одной и той же коллекции, построенные на разных словарях.

4.2 Когерентность

Интерпретируемость тематической модели является плохо ормализуемым 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 инормационно-поисковых системах для систематизации и визуализации результатов тематического поиска или категоризации документов. Эмпирически показано, что внутренней метрикой качества, наиболее кореллирующей с интерпретируемостью, является когерентность.

Тема называется когерентной (согласованной), если термы, наиболее

частые в данной теме, неслучайно часто совместно встречаются рядом в документах коллекции. Численной мерой когерентности темы t является поточечная взаимная информация, вычисляемая по k наиболее вероятным словам темы:

$$Coher(t) = \frac{2}{k(k-1)} \sum_{i=1}^{k-1} \sum_{j=1}^{k} PMI(w_i; w_j)$$

где w_i - i-й терм в порядке убывания φ_{wt}, число k обычно полагается равным 10.

4.3 Разреженность матриц

Как говорилось в подсекции 3.2, предполагается, что каждый документ $d \in D$ и каждый термин $w \in W$ связан с небольшим числом тем $t \in T$. Таким образом, большая часть вероятностей φ_{wt} и θ_{td} должна обнуляться.

Разреженность матрицы Φ или Θ тематической модели измеряется долей нулевых элементов. Если множество тем T разделяется на предметные S и фоновые G так, что $T = S \sqcup G$, то разреженности могут считаться по соответственным разбиениям S и G независимо.

4.4 Характеристики ядер тем

Предполагается, что каждая тема должна содержать своё лексическое ядро, т.е. множество слов, существенно отличающее текущую тему от остальных. Формально ядро W_t темы t определяется как множество терминов, которые имеют высокую условную вероятность $\mathsf{P}(t\,|\,w) = \varphi_{wt}\frac{n_t}{n_w}$ для данной темы $W_t = \{w \in W | \mathsf{P}(t\,|\,w) > threashold\}$.

По ядру определяются следующие показатели интерпретируемости темы t:

$\text{purity}_\text{t} = \sum\limits_{w \in W_t} \mathsf{P}(w\,|\,t)$ — чистота темы (чем выше, тем лучше);

$\text{contrast}_\text{t} = \frac{1}{|W_t|} \sum\limits_{w \in W_t} \mathsf{P}(t\,|\,w)$ — контрастность темы (чем выше, тем

лучше);

kernel_t size = $|W_t|$ — размер ядра (ориентировочный оптимум $\frac{|W|}{|T|}$).

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

5 Постановка задачи

5.1 Оффлайн алгоритм

Оптимизируемым функционалом является линейная комбинация логарифма правдоподобия $L(\Phi, \Theta)$ и r регуляризаторов $R_i(\Phi, \Theta)$ с коэффициентами τ_i:

$$R(\Phi, \Theta) = \sum_{i=1}^{r} \vec{\tau}_i R_i(\Phi), \quad L(\Phi, \Theta) + R(\Phi, \Theta) \to \max_{\tau, \Phi, \Theta}. \tag{5.1}$$

Для решения данной задачи оптимизации применяется итерационный процесс, порождающий последовательность приближений $\{\Phi, \Theta\}^l,\ l = 0,\ 1, \ldots$, где используются правила $\{\Phi, \Theta\}^{l+1} = F_l(\vec{\tau}^l, D)$ с вектором коэффициентов регуляризации $\vec{\tau}^l := \{\tau_i^l\}_{i=1}^n$. Последовательность векторов $\vec{\tau}^l$, $l = 0, 1, \ldots$, называется *траекторией регуляризации*.

Первой задачей текущей работы является адаптивный подбор вектора коэффициентов регуляризации $\vec{\tau}$ при итерировании по коллекции документов D, чтобы как можно быстрее найти такое значение $\vec{\tau}^*$, при котором получится точка оптимума $\{\Phi, \Theta\}^*$ функционала $L(\Phi, \Theta) \to \max$, имеющая желаемые значения показателей качества $\vec{Q}^*$.

На каждой итерации l фиксируется вектор коэффициентов регуляризации $\vec{\tau}^l$ и несколько проходов ЕМ-алгоритма по все коллекции документов. По входным данным и текущим приближениям матриц Φ и Θ вычисляется вектор показателей качества $\vec{Q}^l$.

Ниже приведён псевдокод алгоритма 3, который адаптивно управляет траекторией регуляризации.

Algorithm 3 Адаптивное управление траекторией регуляризации

Вход: коллекция документов $D \equiv \{D_b\}_{b=1}^{B}$, $\vec{\tau}^0$;
Выход: оптимальный вектор коэффициентов регуляризации $\vec{\tau}^*$;

1: **для** итераций $l = 0, 1, \ldots$:
2: $\quad \{\Phi, \Theta\}^l := F_l(\vec{\tau}^l, D)$;
3: $\quad \vec{Q}^l := \vec{Q}(\{\Phi, \Theta\}^l)$;
4: $\quad \vec{\tau}^{l+1} :=$ *выбрать следующий* $\vec{\tau}(\vec{Q}^l, \vec{\tau}^l, \{\Phi, \Theta\}^l, D)$;
5: **end для**

5.2 Онлайн алгоритм

Оптимизируемым функционалом является линейная комбинация логарифма правдоподобия $L(\Phi, \Theta)$ и r регуляризаторов $R_i(\Phi, \Theta)$ с коэффициентами τ_i:

$$R(\Phi) = \sum_{i=1}^{r} \vec{\tau}_i R_i(\Phi), \quad L(\Phi) + R(\Phi) \to \max_{\Phi}. \tag{5.2}$$

Точкой этого функционала является только матрица Φ, основываясь на которой вычисляется матрица Θ при обработке очередной группы документов D_b используя *алгоритм* **??** *обработки пакета документов* (D_b, Φ).

Для решения данной оптимизационной задачи (5.2) применяется итерационный процесс, порождающий последовательность приближений Φ^l, $l = 0, 1, \ldots$, где используются правила $\Phi^{l+1} = F_l(\Phi^l, \vec{\tau}^l, D_{b(l)})$ с вектором коэффициентов регуляризации $\vec{\tau}^l := \{\tau_i^l\}_{i=1}^{n}$. Последовательность векторов $\vec{\tau}^l$, $l = 0, 1, \ldots$, называется *траекторией регуляризации.*

Второй задачей данной работы является адаптивный подбор вектора коэффициентов регуляризации $\vec{\tau}$ при итерировании по коллекции документов D, чтобы как можно быстрее найти такое значение $\vec{\tau}^*$, при котором получится точка оптимума Φ^* функционала $L(\Phi) \to \max$, имеющая желаемые значения показателей качества $\vec{Q}^*$.

На каждой итерации обработки пачки документов l фиксируется вектор коэффициентов регуляризации $\vec{\tau}^l$ и обрабатывается одна пачка $D_{b(l)}$.

По данным этой пачки и текущим приближениям матриц Φ и Θ вычисляется вектор показателей качества $\vec{Q}^l$.

Ниже приведён псевдокод алгоритма 4, который адаптивно управляет траекторией регуляризации.

Algorithm 4 Адаптивное управление траекторией регуляризации

Вход: коллекция документов $D \equiv \{D_b\}_{b=1}^B$, Φ^0, $\vec{\tau}^0$;
Выход: оптимальный вектор коэффициентов регуляризации $\vec{\tau}^*$;

1: **для** итераций $l = 0, 1, \ldots$:
2: $D_{b(l)}$:= *взять следующую пачку из* $\{D_b\}_{b=1}^B$;
3: $\Phi^{l+1} := F_l(\Phi^l, \vec{\tau}^l, D_{b(l)})$;
4: $\vec{Q}^{l+1} := \vec{Q}(x^{l+1})$;
5: $\vec{\tau}^{l+1}$:= *выбрать следующий* $\vec{\tau}(\vec{Q}^{l+1}, \vec{\tau}^l, \Phi^{l+1}, D_{b(l)})$;
6: **end для**

Далее, текущие батчи документов D_b предлагается объединять в группы D_p и обрабатывать в рамках одной итерации l, фиксируя текущий вектор коэффициентов регуляризации $\vec{\tau}^l$. Таким образом, алгоритм 4 теперь дополнительно делает разбиение коллекции на группы пачек $D \equiv \{D_p\}_{p=1}^P$, на которых будут происходить итерации метода, где каждая группа состоит из пачек документов $D_p \equiv \{D_b\}_{b \in B_p}$, $B > P$, не зависящих от текущей итерации l.

Такое объединение пакетов документов в группы D_p позволяет использовать в качестве функции $F_l(\Phi^l, \vec{\tau}^l, D_{b(l)})$ *Online EM-алгоритм* **??** *для модели APTM*, в котором возможны изменения вектора коэффициентов регуляризации в течении работы Online EM-алгоритма.

Подобное объединение документов в группы пакетов уточняет значения показателей качества $\vec{Q}$ и улучшает сходимость APTM моделей.Вдобавок, предложенный алгоритм управления траекторией регуляризации возможно реализовать как надстройку над существующими Online EM-алгоритмами.

Ниже приведён алгоритм 5, дополнительно группирующий пачки документов для лучшей сходимости.

Algorithm 5 Адаптивное управление траекторией регуляризации

Вход: коллекция документов $D \equiv \{D_b\}_{b=1}^{B}$, Φ^0, $\vec{\tau}^0$, размер группы π;
Выход: оптимальный вектор коэффициентов регуляризации $\vec{\tau}^*$;

1: **для** итераций $l = 0, 1, \ldots$:
2: $D_p :=$ *взять следующую группу пачек* $(\{D_b\}_{b=1}^{B}, \pi)$;
3: $\Phi^{l+1} :=$ *Online EM-алгоритм для модели* $APTM(D_p, \vec{\tau}^l)$;
4: $\vec{Q}^{l+1} := \vec{Q}(\Phi^{l+1})$;
5: $\vec{\tau}^{l+1} :=$ *выбрать следующий* $\vec{\tau}(\vec{Q}^{l+1}, \vec{\tau}^l, \Phi^{l+1}, D_p)$;
6: **end для**

Таким образом, остаётся разобраться с методом *выбора следующего вектора коэффициентов регуляризации* $\vec{\tau}$. При постороении траетории регуляризации существуют следующие сложности.

Во-первых, текущие показатели качества Q^l могут быть лишь косвенным образом связаны с регуляризаторами R_i.

Во-вторых, исходная функция $a : \vec{\tau} \to \vec{Q}$ является сложновычислимой и недифференцируемой из-за сложности регуляризаторов, что ограничивает класс методов решения этой задачи методами оптимизации для оракула нулевого порядка.

Кроме того, коэффициенты регуляризации $\vec{\tau}_i$ противоречить друг другу и оптимизировать их поотдельности не представляется возможным.

Часть II

Обзор методов выбора траектории регуляризации

Основываясь на результатах работы [2], в которой было исследовано адаптивное построение пути регуляризации на коллекции синтетических данных, для построения траектории регуляризации в рамках данной работы предлагается рассмотреть несколько методов *выбора следующей точки вектора регуляризации* $\vec{\tau}$, которые будут использоваться в алгоритме 3 и в алгоритме 4. Такие методы, как случайный поиск и жадный поиск с экспертным выбором следующей точки, сейчас используются на практике. При этом итерации алгоритма **??** используют не группы пачек документов D_p, а полностью всю коллекцию D на каждой итерации, что существенно увеличивает время нахождения целевого вектора коэффициентов регуляризации $\vec{\tau}^*$ и затрудняет использование данного метода на потоковых данных, однако осмысленно на небольших коллекциях.

Помимо упомянутых выше методов будет рассматриваться случайный поиск, симплекс метод Нелдера—Мида [54] и Tree-Structured Parzen Estimators [55] в качестве базовых решений, которые хорошо зарекомендовали себя в задачах оптимизации функций с оракулом нулевого порядка.

В данной работе не будут рассматриваться такие популярные оптимизационные методы, как градиентный спуск, метод Ньютона, квазиньютоновские методы, метод сопряжённых градиентов, BFGS, L-BFGS, т.к. они требуют дифференцируемости оптимизируемого функционала, которой в текущем случае может не быть из-за произвольности добавляемых регуляризаторов. Кроме того, не рассматриваются и субградиентные методы, т.к. они требуют выпуклости, которой тоже может не быть.

6 Случайный поиск

Метод случайного поиска предполагает случайный выбор вектора коэффициентов регуляризации $\vec{\tau}$ из фиксированной сетки, которая задаётся экспертно. Хорошего качества или какой-либо сходимости от такого метода ждать не стоит даже при идеальной сетке, но он даёт отличный ориентир на то, какие отклонения может иметь выбранный вектор $\vec{\tau}$ от целевого значения $\vec{\tau}^*$ в выбранных диапазонах сетки.

7 Симплекс метод Nelder—Mead

Симплекс метод Нелдера—Мида [54] — это популярный метод безусловной оптимизации, который не использует градиентов функции и ограничивается только значениями, благодаря чему может применяться к текущей негладкой и зашумлённой задаче. Симплекс метод находит локальный экстремум.

Текущий метод заключается в построении симплекса из $n + 1$ точки в n-мерном пространстве значений. Далее текущий симплекс предлагается итеративно деформировать и перемещать в поисках экстремума путём замены одной из точек. Деформации и перемещения обеспечиваются функциями отражения, сжатия и растяжения.

Для сравнения с другими методами была рассмотрена реализация симплекс метода *minimize(method='Nelder-Mead')* из библиотеки *scipy* (`https://www.scipy.org`) для языка программирования *Python* (`https://www.python.org`).

8 Tree-Structured Parzen Estimators

TPE [55] принимает на вход иерархическое пространство поиска с априорными вероятностями, и на каждом шаге с помощью метода Парзеновского окна уточняет распределения «хороших» и «плохих» точек, основываясь на значениях целевой метрики. В начале каждой итерации

алгоритма выбирается точка, которая наиболее вероятно будет относиться к группе «хороших» точек и менее вероятно к группе «плохих».

Для сравнения с другими методами была доработана и рассмотрена реализация TPE из библиотеки *hyperopt*.

9 Модификации случайного поиска по сетке

Случайный поиск(см. секцию 6) можно существенно улучшить с помощью множественной регрессии. В этом методе предлагается восстанавливать регрессионную зависимость вектора показателей качества $\vec{Q}$ от вектора параметров $\vec{\tau}$.

В ходе итерационного процесса накапливается обучающая выборка пар векторов $(\vec{\tau}^l, \vec{Q}^l)$, где вектор $\vec{Q}^l$ — это значения показателей качества, полученные при построении APTM модели с коэффициентами регуляризации $\vec{\tau}^l$. Если зависимость $\vec{Q}(\vec{\tau})$ удалось хорошо восстановить, то в качестве следующей точки приближения целевого вектора $\vec{Q}^*$ можно брать наиближайший из $\{\vec{Q}(\vec{\tau}^k)\}_{k=0}^K$, где вектор коэффициентов регуляризации $\vec{\tau}^k$ перебирается по заранее фиксированной сетке или ее случайного подмножества. Если регрессионная модель строит свои предсказания не очень долго по сравнению с разложением APTM модели, то эта операция в данной задаче будет заведомо оправдана.

При восстановлении множественной регрессионной зависимости существуют следующие сложности.

Во-первых, изначально выборка $\{(\vec{\tau}^l, \vec{Q}^l)\}_{l=0}^N$ пуста, и не получится сразу хорошо предсказывать вектор показателей качества для векторов $\vec{\tau}^k$, $k = 0, 1, \ldots, K$. Основная задача состоит в том, чтобы как можно раньше направить *исследовательские* шаги в область пространства $\vec{Q}^*$. Для простоты, можно взять первые N_0 значений случайными.

Во-вторых, достижение окрестности $\vec{Q}^*$ может оказаться трудной задачей, а регрессионная зависимость может строиться по обучающей вы-

борке, которая находится далеко от целевых значений $(\vec{\tau}^*, \vec{Q}^*)$. Таким образом, аппроксимация регрессионной моделью может давать недостаточную точность. Для решения этой проблемы предлагается чередовать *исследовательские* и *максимизирующие* шаги, чтобы исследовать ещё непокрытые регрессионной зависимостью участки и, в дальнейшем, лучше предсказывать целевой вектор. На этапе исследования предлагается выбирать произвольный вектор $\vec{\tau}^l$, а на этапе максимизации делать попытку приблизиться к оптимуму.

10 Модификации TPE

Предыдущую модификацию случайного поиска можно улучшить, если для поиска предполагаемых значений вектора коэффициентов регуляризации $\vec{\tau}^k$, которые мы ранжируем на основании предсказаний множественной регрессии, использовать не перебор по заранее фиксированной сетке, а семплировать «хорошие» точки из распределения TPE.

Во время исследовательского шага можно также брать не случайную точку из пространства возможных значений коэффициентов регуляризации, а наилучшую с точки зрения TPE.

Часть III

Эксперименты

В данной части предлагается рассмотреть описанные выше методы на практике и сравнить их по качеству на реальных данных. У текущих решений есть гиперпараметры и параметры, про которые важно помнить и которые необходимо правильно задавать.

11 Данные

Все эксперименты проводились на 2 коллекциях документов:

- **postnauka** - postnauka.ru
  $|\mathrm{D}| = 3446$ $|\mathrm{W}| = 35531$;

- **elementy** - elementy.ru
  $|\mathrm{D}| = 2397$ $|\mathrm{W}| = 54349$;

Количество тем было взято $|T| = 20$, из которых было $|S| = 19$ предметных тем и $|G| = 1$ фоновая.

Предметные темы $t \in S$ должны содержать термины предметных областей. Чтобы эти темы были интерпретируемыми, они должны содержать понятные по смыслу лексические ядра (см. подсекцию 4.4), т.е. множества слов, характерных для определённых предметных областей, которые часто употребляются рядом в документах, с большой вероятностью употребляются в данных темах и практически не употребляются в других темах. Распределения слов в предметных темах $\mathsf{P}(w\,|\,t)$ должны быть разрежены и существенно различны, т.е. декоррелированы. Распределения документов в предметных темах $\mathsf{P}(d\,|\,t)$ также должны быть разрежены, так как каждая предметная тема должна присутствовать в относительно небольшой доле документов, а в каждом документе должно обсуждаться не более пары тем.

Фоновые темы $t \in G$ должны содержать слова общей лексики, которых не должно быть в предметных темах. Их распределения $\mathsf{P}(w|t)$ и $\mathsf{P}(d|t)$ сглажены, так как эти слова присутствуют в большинстве документов.

12 Регуляризаторы и метрики качества

В текущих экспериментах будем использовать рассмотренные выше регуляризаторы (см. секцию 3) и метрики качества (см. секцию 4).

13 Вектор целевых значений функционалов качества $\vec{Q}^*$ и метрика близости

Текущий метод адаптивного подбора траектории регуляризации пытается найти такую точку в пространстве коэффициентов регуляризации, которая приближает значения некоторого функционала качества к целевому его значению. Эталонные значения вектора функционала качества задаются наилучшими теоретическими оценками метрик качества (см. секцию 4).

Для измерения степени близости до целевого значения вектора функционалов качества необходимо задать метрику. В текущих экспериментах будет измеряться взвешенное отклонение *(weighted error)* от целевого вектора $\vec{Q}^*$:

$$wE(\vec{Q}, \vec{Q}^*) = \frac{\sum_{j=0}^{r} w_j(\vec{Q}_j - \vec{Q}_j^*)}{m}; \tag{13.1}$$

Веса нужны для приведения значений показателей качества к единой шкале, а их подбор требует непосредственных экспертных знаний об относительной допустимости отклонений. Таким образом, веса имеют естественную интерпретацию важности сходимости того или иного критерия качества, но заранее не приведены к единому масштабу.

14 Итеративное обучение исследовательских методов

Методы из секций 8, 9, 10 требуют дополнительных установок и уточнений по начальным и основным итерациям. Ниже описано, какие конкретно шаги и модели для исследований использовались в текущих экспериментах.

14.1 Начальные итерации исследовательских методов

В методах секций 8, 9, 10 в текущих экспериментах в качестве начальной инициализации брались значения «базисных» векторов. В качестве первого «базисного» вектора был взят нулевой вектор $\vec{0}$, а остальные «базисные» векторы брались путём изменения одной из координат нулевого вектора, соответствующей конкретному регуляризатору. Направления, в которых нужно изменять регуляризаторы, задаются экспертно, но если такой информации нет, то можно шагать по каждой координате в обе стороны.

Таким образом, получается $m+1$ пара векторов $(\vec{\tau}, \vec{Q})$, где m — это число используемых регуляризаторов для обучения APTM модели. Эти пары в дальнейшем идут на обучение первых моделей для восстановления регрессионной зависимости.

14.2 Основные итерации исследовательских методов

На исследовательских основных шагах в секции 9 предлагалось брать случайную точку, а в секции 10 тот вектор, для которого отношение вероятностей быть «хорошим» и быть «плохим» наиболее велико. Все эти подходы будут использованы и сравнены в секции 15.

На на основных шагах максимизирования будет браться точка, по-

лучившая лучшее предсказание модели градиентного бустинга.

15 Сравнение результатов

Ниже приведены результаты сходимости методов к целевому вектору значений показателей качества $\vec{Q}^*$ по метрике wE, рассмотренной в секции 13.

15.1 Оффлайн обучение

15.1.1 Постнаука

На рисунке 1 представлены результаты базовых алгоритмов — случайный поиск (см. секцию 6) и симплекс метод Nelder—Mead (см. секцию 7). Они дают базовые ориентиры на понимание принимаемых значений метрики wE и оценки скорости сходимости методов.

Рис. 1: Сравнение случайного поиска и Нельдер-Мида

На рисунках 2, 3 рассмотрены графики сходимостей модифицированного случайного поиска(см. секцию 9) и его улучшения с использованием

метода TPE (см. секцию 8). Пунктирными линиями показаны исследовательские шаги (exploration), которые могут принимать произвольные значения метрики, а сплошными линиями показаны максимизирующие шаги (maximization), на которые и стоит ориентироваться, сравнивая сходимость методов.

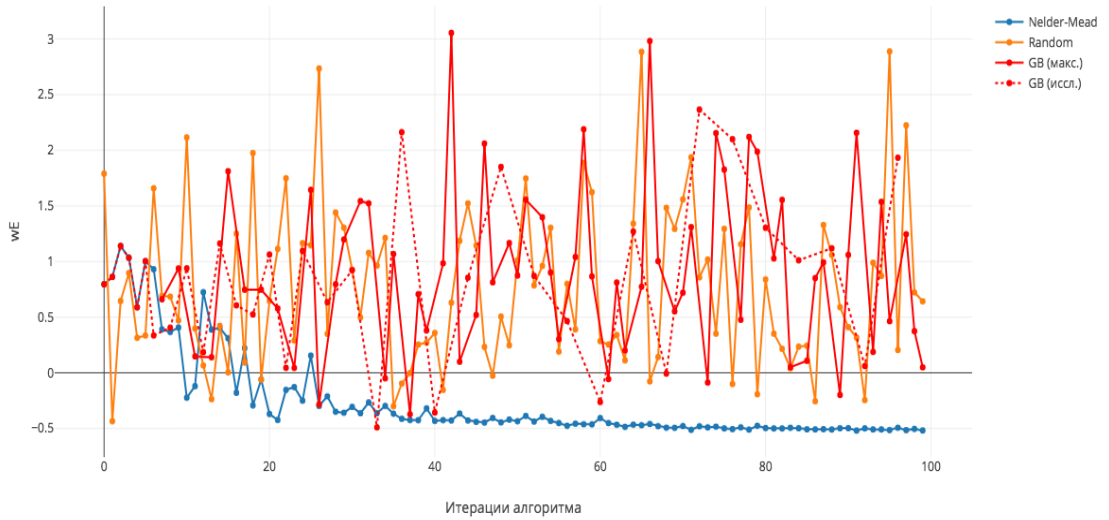

Рис. 2: Сравнение случайного поиска, Нельдер-Мида и модифицированного случайного посика

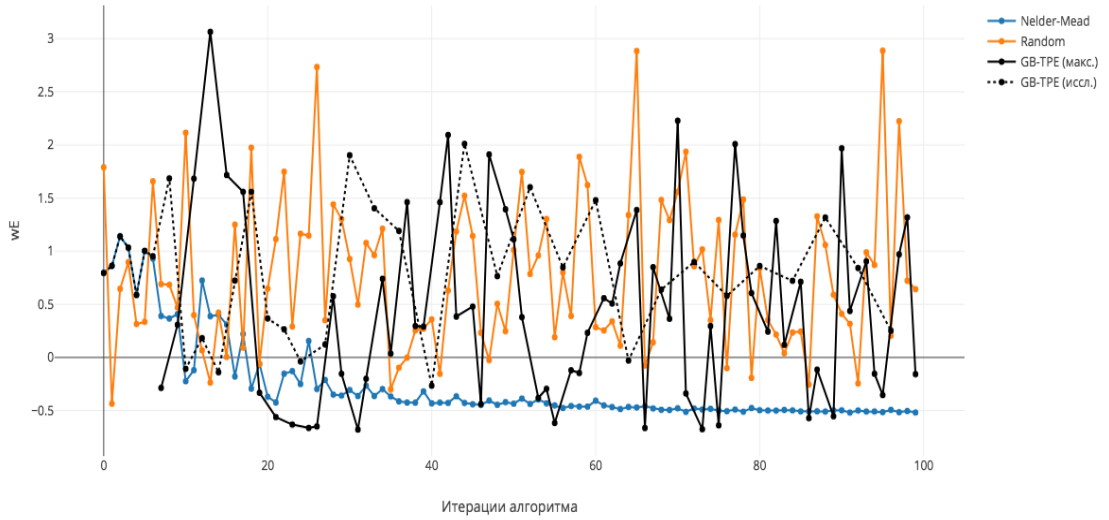

Рис. 3: Сравнение случайного поиска, Нельдер-Мида и модифицированного TPE

15.1.2 Элементы

На рисунке 4 представлены результаты базовых алгоритмов — случайный поиск (см. секцию 6) и симплекс метод Nelder—Mead (см. секцию 7). Они дают базовые ориентиры на понимание принимаемых значений метрики wE и оценки скорости сходимости методов.

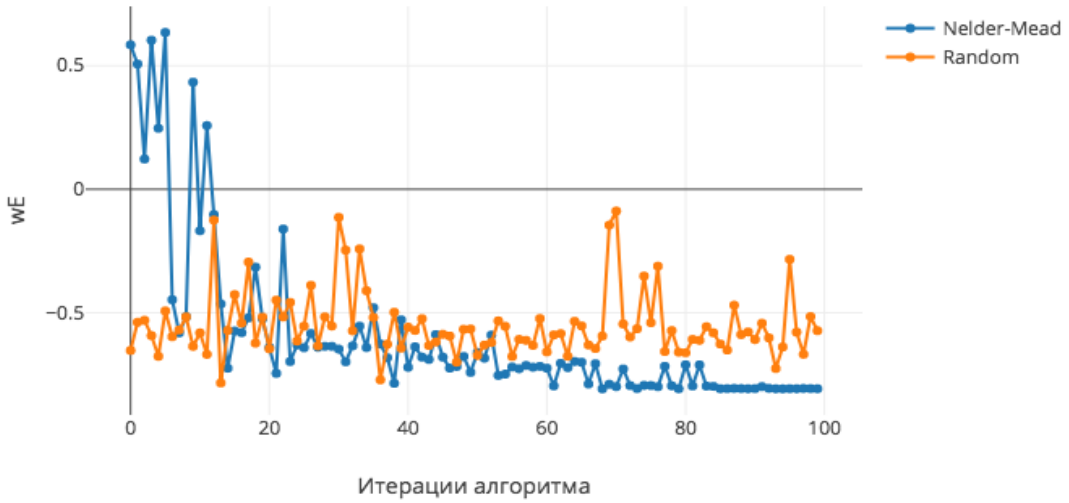

Рис. 4: Сравнение случайного поиска и Нельдер-Мида

На рисунках 5, 6 рассмотрены графики сходимостей модифицированного случайного поиска(см. секцию 9) и его улучшения с использованием метода TPE (см. секцию 8). Пунктирными линиями показаны исследовательские шаги (exploration), которые могут принимать произвольные значения метрики, а сплошными линиями показаны максимизирующие шаги (maximization), на которые и стоит ориентироваться, сравнивая сходимость методов.

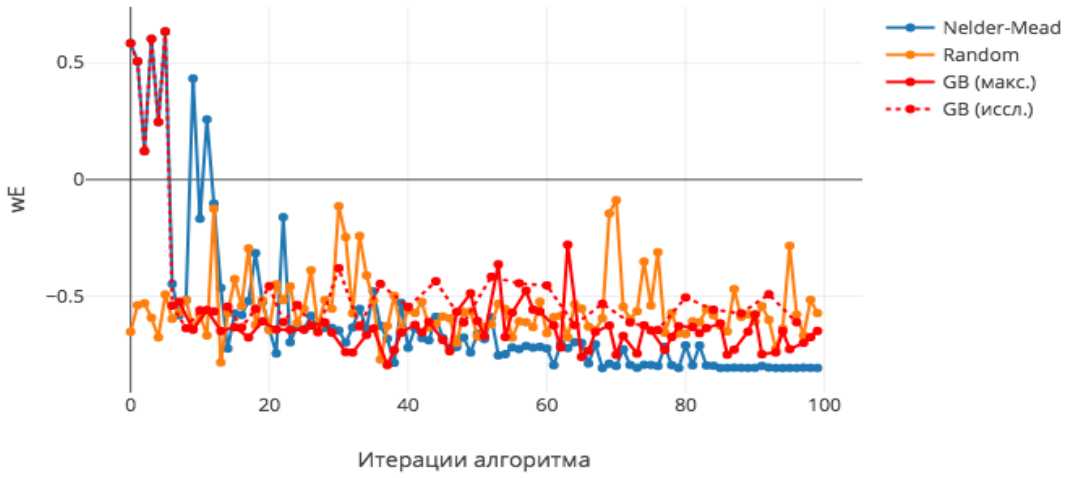

Рис. 5: Сравнение случайного поиска, Нельдер-Мида и модифицированного случайного посика

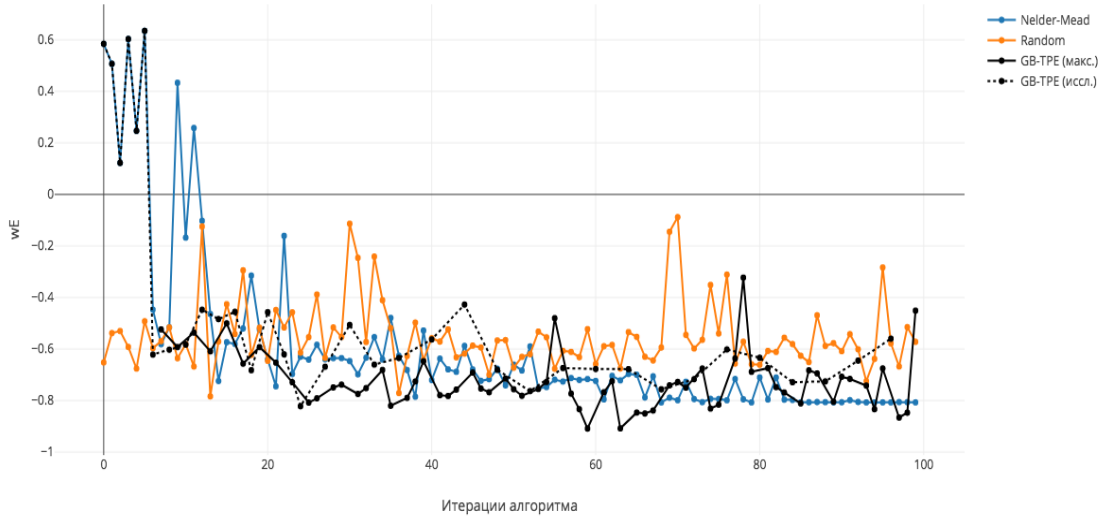

Рис. 6: Сравнение случайного поиска, Нельдер-Мида и модифицированного TPE

15.2 Онлайн обучение

Для онлайн эксперимента были выбраны следующие параметры для разбиения на группы пакетов документов:

- размер пакета документов - 20

- количество групп - 8

- в среднем 20 пакетов в группе

15.2.1 Постнаука

На рисунке 7 представлены результаты базовых алгоритмов — случайный поиск (см. секцию 6) и симплекс метод Nelder—Mead (см. секцию 7). Они дают базовые ориентиры на понимание принимаемых значений метрики wE и оценки скорости сходимости методов.

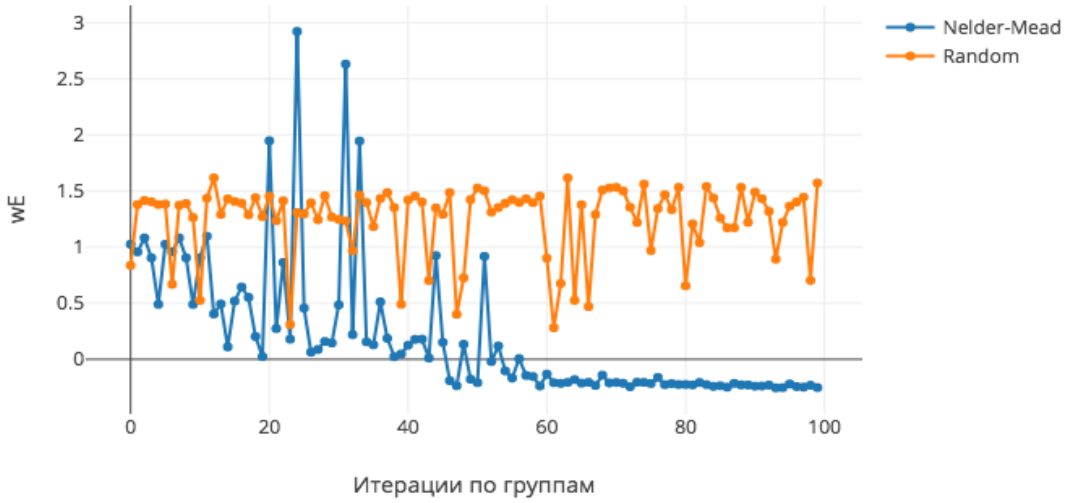

Рис. 7: Сравнение случайного поиска и Нельдер-Мида

На рисунках 8, 9 рассмотрены графики сходимостей модифицированного случайного поиска(см. секцию 9) и его улучшения с использованием метода TPE (см. секцию 8). Пунктирными линиями показаны исследовательские шаги (exploration), которые могут принимать произвольные значения метрики, а сплошными линиями показаны максимизирующие шаги (maximization), на которые и стоит ориентироваться, сравнивая сходимость методов.

Рис. 8: Сравнение случайного поиска, Нельдер-Мида и модифицированного случайного посика

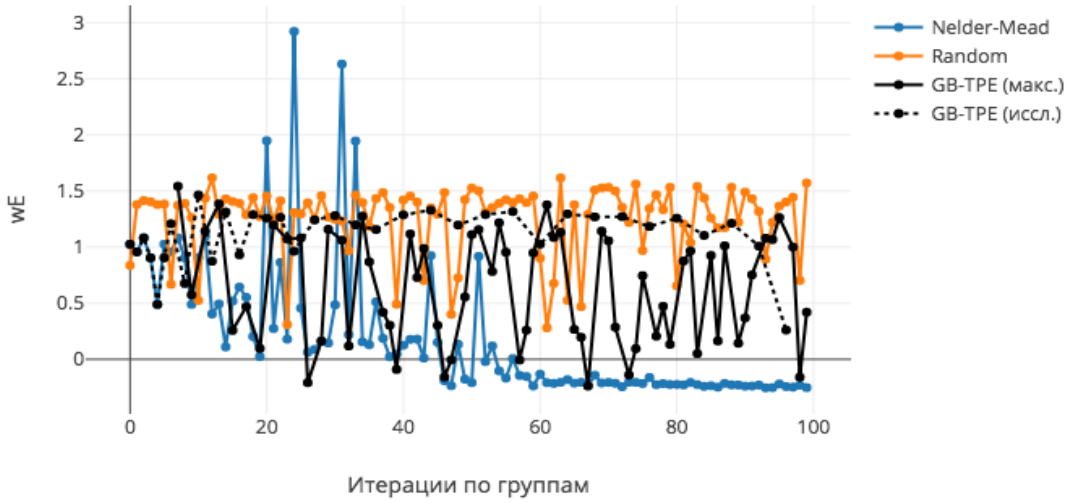

Рис. 9: Сравнение случайного поиска, Нельдер-Мида и модифицированного TPE

15.2.2 Элементы

На рисунке 10 представлены результаты базовых алгоритмов — случайный поиск (см. секцию 6) и симплекс метод Nelder—Mead (см. сек-

цию 7). Они дают базовые ориентиры на понимание принимаемых значений метрики wE и оценки скорости сходимости методов.

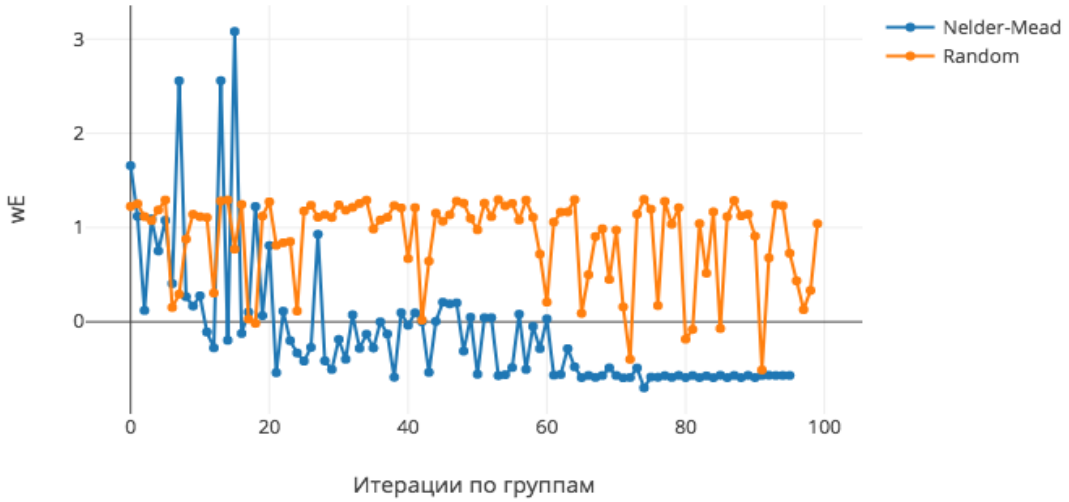

Рис. 10: Сравнение случайного поиска и Нельдер-Мида

На рисунках 11, 12 рассмотрены графики сходимостей модифицированного случайного поиска(см. секцию 9) и его улучшения с использованием метода TPE (см. секцию 8). Пунктирными линиями показаны исследовательские шаги (exploration), которые могут принимать произвольные значения метрики, а сплошными линиями показаны максимизирующие шаги (maximization), на которые и стоит ориентироваться, сравнивая сходимость методов.

Рис. 11: Сравнение случайного поиска, Нельдер-Мида и модифицированного случайного посика

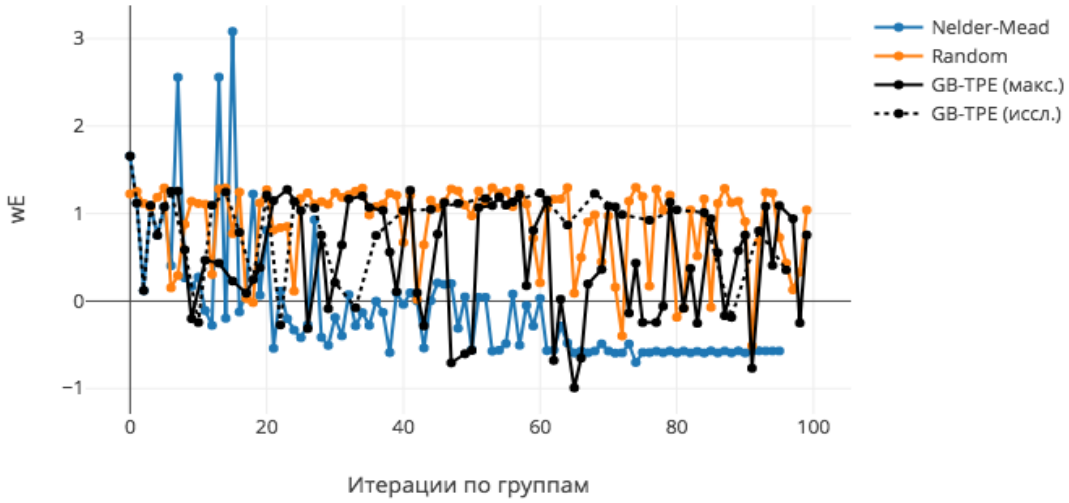

Рис. 12: Сравнение случайного поиска, Нельдер-Мида и модифицированного TPE

15.3 Интерпретация графиков

Пики резкого ухудшения метрик обычно возникают из-за того, что один из показателей качества Q_j сходится к плохому значению, что может возникать при небольших отступах от предыдущих «хороших» значений вектора коэффициентов регуляризации $\vec{\tau}^l$.

Из рисунков видно, что модифицированные методы с использование оценки TPE показывают более хорошее качество по сравнению с симплекс методом и модифицированными методами случайного поиска, которые, в свою очередь, лучше обычного случайного поиска. Начиная с некоторой итерации, симплекс метод выходит на плато, скатываясь в одну не меняющуюся точку. При этом, модифицированный метод с использование оценки TPE на каждой итерации выбирает случайную, отличную от предыдущих, точку, переодически опускаясь ниже Нельдер-Мида с точки зрения интересующей метрики.

Часть IV

Выводы

Таким образом, в данной работе разработаны новые эффективные методы адаптивного подбора траектории регуляризации в тематическом моделировании (см. часть II), которые естественным образом обобщают текущие методы подбора коэффициентов регуляризации[2].

Кроме того, проведены эксперименты по сравнению различных методов выбора траектории регуляризации на реальных данных (см. часть III), в результате чего отмечается, что лучшими с точки зрения метрики сходимостями при решении текущей задачи обладают модификация случайного поиска с использованием TPE (см. секции 8 и алгоритм Нельдера-Мида 7).

Можно увидеть, что предлагаемое решение обладает как положительными свойствами, так и отрицательными.

К положительным свойствам можно отнести автоматизацию подбора траектории регуляризации и отсутствие необходимости экспертных знаний для его запуска. Кроме того, предложенй способ позволяет находить коэффициенты регуляризации как при оффлайн, так и при онлайн обучении. Разработанные методы допускают подбор коэффициентов для произвольных регуляризаторов и могут настраиваться на различные функционалы качества, которые могут не обладать свойствами дифференцируемости и гладкости.

Предложенный метод обладает и недостатками. К отрицательным свойствам текущего подхода можно отнести то, что для эксплуатации требуется многократный запуск итераций алгоритма поиска. Однако, как видно из графиков, практически во всех случаях, алгоритм модификация случайного поиска с использованием TPE (см. секции 8) находил точку, близкую к оптимальной, уже на 20-30 итерации.

Список литературы

[1] Воронцов К. В. Вероятностное тематическое моделирование — 2013.

[2] Кузьмин А. Н. Адаптивный выбор траектории регуляризации — 2017.

Daud A., Li J., Zhou L., Muhammad F. Knowledge discovery through directed probabilistic topic models: a survey // Frontiers of Computer Science in China. — 2010.

[3] Zhang J., Song Y., Zhang C., Liu S. Evolutionary hierarchical Dirichlet processes for multiple correlated time-varying corpora // Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. — 2010.

[4] Cui W., Liu S., Tan L., Shi C., Song Y., Gao Z., Qu H., Tong X. TextFlow: Towards better understanding of evolving topics in text. IEEE transactions on visualization and computer graphics. — 2011.

[5] Rubin T. N., Chambers A., Smyth P., Steyvers M. Statistical topic models for multilabel document classification // Machine Learning. — 2012.

[6] Zhou S., Li K., Liu Y. Text categorization based on topic model // International Journal of Computational Intelligence Systems. — 2009.

[7] Hospedales T., Gong S., Xiang T. Video behaviour mining using a dynamic topic model // International Journal of Computer Vision. — 2012.

[8] Li X.-X., Sun C.-B., Lu P., Wang X.-J., Zhong Y.-X. Simultaneous image classification and annotation based on probabilistic model // The Journal of China Universities of Posts and Telecommunications. — 2012.

[9] Feng Y., Lapata M. Topic models for image annotation and text illustration // Human Language Technologies: The 2010 Annual

Conference of the North American Chapter of the Association for Computational Linguistics. — Association for Computational Linguistics, 2010.

[10] 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.

[11] 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.

[12] Yi X., Allan J. A comparative study of utilizing topic models for information retrieval // Advances in Information Retrieval. — Springer Berlin Heidelberg, 2009.

[13] Andrzejewski D., Buttler D. Latent topic feedback for information retrieval // Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. — KDD 2011.

[14] Steyvers M., Griffiths T. Finding scientific topics // Proceedings of the National Academy of Sciences. — 2004.

[15] Blei D., Lafferty J. A correlated topic model of Science // Annals of Applied Statistics. — 2007.

[16] Bolelli L., Ertekin S., Giles C. L. Topic and trend detection in text collections using latent dirichlet allocation // ECIR. — Springer, 2009.

[17] Airoldi E. M., Erosheva E. A., Fienberg S. E., Joutard C., Love T., Shringarpure S. Reconceptualizing the classification of pnas articles // Proceedings of The National Academy of Sciences. — 2010.

[18] Paul M. J., Girju R. Topic modeling of research fields: An interdisciplinary perspective // RANLP. — 2009 Organising Committee / ACL, 2009.

[19] 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

[20] Vulić I., De Smet W., Tang J., Moens M.-F. Probabilistic topic modeling in multilingual settings: an overview of its methodology and applications // Information Processing & Management. — 2015.

[21] Vulić I., Smet W., Moens M.-F. Cross-language information retrieval models based on latent topic models trained with document-aligned comparable corpora // Information Retrieval. — 2012.

[22] Krestel R., Fankhauser P., Nejdl W. Latent dirichlet allocation for tag recommendation // Proceedings of the third ACM conference on Recommender systems. — ACM, 2009.

[23] Zhao X. W., Wang J., He Y., Nie J.-Y., Li X. Originator or propagator?: Incorporating social role theory into topic models for Twitter content analysis // New York, NY, USA: ACM, 2013.

[24] Varshney D., Kumar S., Gupta V. Modeling information diffusion in social networks using latent topic information // Intelligent Computing Theory / Ed. by D.-S. Huang, V. Bevilacqua, P. Premaratne. — Springer International Publishing, 2014.

[25] Pinto J. C. L., Chahed T. Modeling multi-topic information diffusion in social networks using latent Dirichlet allocation and Hawkes processes // Tenth International Conference on Signal-Image Technology & Internet-Based Systems. — 2014.

[26] Павлов А. С., Добров Б. В. Метод обнаружения массово порожденных неестественных текстов на основе анализа тематической структуры // Вычислительные методы и программирование: новые вычислительные технологии. — 2011.

[27] Yeh J.-h., Wu M.-l. Recommendation based on latent topics and social network analysis // Proceedings of the 2010 Second International Conference on Computer Engineering and Applications. — Vol. 1. — IEEE Computer Society, 2010.

[28] Yin H., Cui B., Chen L., Hu Z., Zhang C. Modeling location-based user rating profiles for personalized recommendation // ACM Transactions of Knowledge Discovery from Data. — 2015.

[29] Yin H., Cui B., Sun Y., Hu Z., Chen L. Lcars: A spatial item recommender system // ACM Transaction on Information Systems. — 2014.

[30] Lee S. S., Chung T., McLeod D. Dynamic item recommendation by topic modeling for social networks // Information Technology: New Generations (ITNG), 2011 Eighth International Conference on. — IEEE, 2011.

[31] La Rosa M., Fiannaca A., Rizzo R., Urso A. Probabilistic topic modeling for the analysis and classification of genomic sequences // BMC Bioinformatics. — 2015.

[32] Shivashankar S., Srivathsan S., Ravindran B., Tendulkar A. V. Multi-view methods for protein structure comparison using latent dirichlet allocation. // Bioinformatics [ISMB/ECCB]. — 2011.

[33] Konietzny S., Dietz L., McHardy A. Inferring functional modules of protein families with probabilistic topic models // BMC Bioinformatics. — 2011.

[34] Pritchard J. K., Stephens M., Donnelly P. Inference of population structure using multilocus genotype data // Genetics. — 2000.

[35] Воронцов К. В., Потапенко А. А. Аддитивная регуляризация тематических моделей — 2014.

[36] Воронцов К. В., Потапенко А. А. Регуляризация вероятностных тематических моделей для повышения интерпретируемости и определения числа тем // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 4–8 июня 2014 г.). — Вып. 13 (20). — М: Изд-во РГГУ, 2014.

[37] Vorontsov K. V., Potapenko A. A. Additive regularization of topic models // Machine Learning, Special Issue on Data Analysis and Intelligent Optimization. — 2014.

[38] Vorontsov K. V., Potapenko A. A. Tutorial on probabilistic topic modeling: Additive regularization for stochastic matrix factorization // AIST'2014, Analysis of Images, Social networks and Texts. — Springer International Publishing Switzerland, Communications in Computer and Information Science (CCIS), 2014.

[39] Hofmann T., Probabilistic Latent Semantic Indexing, Proceedings of the Twenty-Second Annual International SIGIR Conference on Research and Development in Information Retrieval (SIGIR-99) — 1999.

[40] Dempster A. P., Laird N. M., Rubin D. B. Maximum likelihood from incomplete data via the EM algorithm // J. of the Royal Statistical Society, Series B. — 1977.

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

[42] Bassiou N., Kotropoulos C. Online plsa: Batch updating techniques including out-of-vocabulary words // Neural Networks and Learning Systems, IEEE Transactions on. — Nov 2014.

[43] Hoffman M. D., Blei D. M., Bach F. R. Online learning for latent dirichlet allocation // NIPS. — Curran Associates, Inc., 2010.

[44] Воронцов К. В., Фрей А. И., Апишев М. А., Ромов П. А., Янина А. О., Суворова М. А. BigARTM: библиотека с открытым кодом для тематического моделирования больших текстовых коллекций // Аналитика и управление данными в областях с интенсивным использованием данных. XVII Международная конференция DAMDID/RCDL'2015, Обнинск, 13-16 октября 2015.

[45] Vorontsov K. V., Frei O. I., Apishev M. A., Romov P. A., Suvorova M. A. BigARTM: Open Source Library for Regularized Multimodal Topic Modeling of Large Collections // AIST'2015, Analysis of Images, Social networks and Texts. Springer International Publishing Switzerland, 2015. Communications in Computer and Information Science (CCIS).

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

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

[48] Khalifa O., Corne D., Chantler M., Halley F. Multi-objective topic modelling // 7th International Conference Evolutionary Multi-Criterion Optimization (EMO 2013). — Springer LNCS, 2013.

[49] 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.

[50] Blei D. M. Probabilistic topic models // Communications of the ACM. — 2012.

[51] Blei D. M., Ng A. Y., Jordan M. I. Latent Dirichlet allocation // Journal of Machine Learning Research. — 2003.

[52] Воронцов К. В., Фрей А. И., Апишев М. А., Потапенко А. А. Тематическое моделирование в BigARTM: теория, алгоритмы, приложения. — 2015.

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

[54] Nelder J. A., Mead R. A simplex method for function minimization. Computer Journa, — 1965.

[55] Bergstra, J., Yamins, D., Cox, D. D. (2013) Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures. To appear in Proc. of the 30th International Conference on Machine Learning (ICML 2013).