МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ

ГОСУДАРСТВЕННОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ
“МОСКОВСКИЙ ФИЗИКО-ТЕХНИЧЕСКИЙ ИНСТИТУТ
(ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ)”

ФАКУЛЬТЕТ ИННОВАЦИЙ И ВЫСОКИХ ТЕХНОЛОГИЙ
КАФЕДРА АНАЛИЗА ДАННЫХ

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ ПО НАПРАВЛЕНИЮ
01.04.02 “ПРИКЛАДНЫЕ МАТЕМАТИКА И ИНФОРМАТИКА”
НА ТЕМУ:

АДАПТИВНЫЙ ВЫБОР ТРАЕКТОРИИ РЕГУЛЯРИЗАЦИИ

Студент _____ А.Н.Кузьмин

Научный руководитель д.ф-м.н _____ К. В. Воронцов

Зав. кафедрой д.ф-м.н _____ Е. И. Бунина

МОСКВА 2017

Содержание

Часть I
Введение

Одной из современных и активно развивающихся областей статистического анализа текстов является *тематическое моделирование* (topic modeling) [1, 34]. Вероятностная тематическая модель выявляет тематику коллекции текстовых документов, описывая каждую тему дискретным распределением на множестве терминов, каждый документ — дискретным распределением на множестве тем. Оно может учитывать различные особенности языка и текстовых коллекций. Тематические модели используются для анализа различных книг, статей, блогов, писем, веб-страниц и др., уже хорошо зарекомендовали себя и используются для выявления основных трендов в научных публикациях или новостных потоках [2, 3], для классификации и категоризации [4, 5] документов, изображений и видеопотоков [6, 7, 8, 9], для тематической сегментации текстов [30], для информационного поиска [10, 11, 23, 24, 25, 26, 27, 28], в том числе многоязычного [29, 12], для тегирования веб-страниц [13], для анализа данных социальных сетей [31, 32, 33], для обнаружения текстового спама [14], для рекомендательных систем [15, 30, 16, 17, 18], для анализа нуклеотидных [19] и аминокислотных последовательностей [20, 21], в задачах популяционной генетики [22] и других приложений из разных областей.

Естественным улучшением тематической модели для решения практических задач было рассмотрение композитных моделей, совмещающие большое число функциональных требований. В частности, для научного поиска по большим коллекциям публикаций нужны были модели, одновременно иерархические, динамические, n-граммные, разреженные, робастные, мультиязычные и т.д. Учитывать несколько свойств одновременно позволила теория аддитивной регуляризации тематических моделей (АРТМ) [38, 39, 40, 41]. Таким образом, появились модели, выявляющие ключевые фразы (термины предметной области), учитывающие морфологию слов и синтаксическую структуру предложений, отслеживающие изменения тематики во времени или внутри отдельных документов, строящие иерархические отношения между темами, учитывающие связи между документами через авторство или ссылки и т.д. Но помимо правдоподобия теперь потребовалось максимизировать ещё n регуляризаторов качества R_i. Для решения, эту многокритериальную задачу обычно сводят к однокритериальной, путём использования линейной комбинации регуляризаторов R_i, но при этом возникает проблема выбора вектора коэффициентов регуляризации $\vec{\tau}$. К настоящему моменту уже накопилось действительно большое число разнообразных регуляризаторов, и они нетривиальным образом влияют друг на друга, поэтому подбор траектории регуляризации стал отдельной сложной задачей.

Целью данной работы является создание метода, который позволял бы эффективно, в процессе обработки коллекции документов, подбирать такой вектор коэффициентов регуляризации $\vec{\tau}^*$, позволяющий получать требуемые значения функционалов качества. Созданный на основе идей из теории оптимальных экспериментов, активного обучения и обучения с подкреплением метод сравнивается с существующими вариантами подбора вектора регуляризации и с популярными методами решения схожей оптимизационной задачи.

Часть II
Вероятностное тематическое моделирование и постановка проблемы

1 Задача тематического моделирования

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.

Найденные распределения используются для решения прикладных задач. Например, распределение $\mathsf{P}(t\,|\,d)$ активно используется для информационного поиска, классификации и категоризации документов.

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:

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

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) 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:

$$F_{WD} \;\approx\; \Phi_{WT} \;\times\; \Theta_{TD}; \tag{1′}$$

$$\Phi_{WT} = (\varphi_{wt})_{W\times T}, \quad \varphi_{wt} = \mathsf{P}(w\,|\,t) = \frac{n_{wt}}{n_t};$$

$$\Theta_{TD} = (\theta_{td})_{T\times D}, \quad \theta_{td} = \mathsf{P}(t\,|\,d) = \frac{n_{td}}{n_d}.$$

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) [35] для построения модели (1′) предлагается максимизировать логарифм правдоподобия (плотности распределения) выборки при ограничениях неотрицательности и нормировки столбцов матриц Φ и Θ:

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

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

Максимум задачи (2) в вероятностном латентном семантическом анализе ищут с помощью итерационного алгоритма EM [36]. 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 условные вероятности всех тем для каждой пары термин-документ:

$$H_{dwt} = \mathsf{P}(t\,|\,d,w) = \frac{\mathsf{P}(w\,|\,t)\,\mathsf{P}(t\,|\,d)}{\mathsf{P}(w\,|\,d)}. \tag{3}$$

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

$$n_{dwt} \approx n_{dw}\,\mathsf{P}(t\,|\,d,w) = n_{dw} H_{dwt}, \tag{4}$$

несложно по условным вероятностям пересчитать новое приближение параметров:

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

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

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

Лагранжиан задачи (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:

$$\frac{\partial \mathscr{L}}{\partial \varphi_{wt}} = \sum_{d \in D} n_{dw} \frac{\theta_{td}}{\mathsf{P}(w \,|\, d)} - \lambda_t = 0. \tag{7}$$

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 (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, который отлично описан в работе [1].

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, 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 [37]. Ниже приведён псевдокод рационального EM-алгоритма 2 для модели PLSA.

Далее, в секции 3 будет рассмотрено, как можно сделать EM-алгоритм онлайновым, что позволяет обрабатывать большие коллекции из-за эффективной схемы

Algorithm 1 EM-алгоритм для модели PLSA

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

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 $\varphi_{wt}\theta_{td} > 0$:
      $\delta := n_{dw}\varphi_{wt}\theta_{td}/Z$;
      $n_{wt} := n_{wt} + \delta$;
      $n_{dt} := n_{dt} + \delta$;
      $n_t := n_t + \delta$;
  $\varphi_{wt} := n_{wt}/n_t$ 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;

вычислений. Именно Online EM-алгоритм [45, 46], а точнее его параллельную реализацию в BigARTM [42, 43], мы и будем использовать в качестве инструмента для постановки экспериментов в части IV.

2 Аддитивная регуляризация тематических моделей

Искомое стохастическое матричное разложение $F \approx \Phi\Theta$ в тематической модели определено не единственным образом. В общем виде задача тематического моделирования имеет бесконечно много решений. Такие задачи принято называть *некорректно поставленными* [48], а общий подход к устранению данной проблемы — *регуляризацией* [49]. Её идея заключается в добавлении к логарифму правдоподобия (2) штрафного слагаемого, которое сужает множество решений.

Построение многоцелевых тематических моделей [53] существенно упрощается благодаря аддитивности регуляризаторов, причём добавление регуляризатора требует лишь небольшой модификации M-шага.

2.1 Постановка и решение задачи

Предположим, что вместе с логарифмом правдоподобия (2) требуется максимизировать ещё r критериев $R_i(\Phi, \Theta)$, $i = 1, \ldots, r$, называемых регуляризаторами. Тогда для оптимизации будем рассматривать линейную комбинацию критериев $L(\Phi, \Theta)$ и $R_i(\Phi, \Theta)$ с неотрицательными *коэффициентами регуляризации* $\hat{\tau}_i$:

$$R(\Phi, \Theta) = \sum_{i=1}^{r} \hat{\tau}_i R_i(\Phi, \Theta), \quad L(\Phi, \Theta) + R(\Phi, \Theta) \to \max_{\Phi, \Theta}. \tag{8}$$

Решение этой задачи приводит к обобщению формул M-шага [38] в EM-

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 $\varphi_{wt}\theta_{td} > 0$:
      $\delta := n_{dw}\varphi_{wt}\theta_{td}/Z$;
      $n_{wt} := n_{wt} + \delta$;
      $n_{dt} := n_{dt} + \delta$;
      $n_t := n_t + \delta$;
  $\varphi_{wt} := n_{wt}/n_t$ 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;

алгоритме:

$$\varphi_{wt} = \frac{\Big(n_{wt} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}}(\Phi,\Theta)\Big)_+}{\sum\limits_{u \in W}\Big(n_{ut} + \varphi_{ut}\dfrac{\partial R}{\partial \varphi_{ut}}(\Phi,\Theta)\Big)_+}, \qquad \theta_{td} = \frac{\Big(n_{dt} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}(\Phi,\Theta)\Big)_+}{\sum\limits_{s \in T}\Big(n_{ds} + \theta_{sd}\dfrac{\partial R}{\partial \theta_{sd}}(\Phi,\Theta)\Big)_+}, \tag{9}$$

где $(x)_+ = \max\{0, x\}$, а значения n_{wt}, n_{dt} определяются формулами (3)–(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 модели PLSA увеличение числа тем может приводить только к росту правдоподобия модели. Для регуляризованной модели это не обязательно так. Модель APTM предусматривает возможность обнуления столбцов матриц Φ и Θ. Если для фиксированной темы t_0 обнуляется столбец $\varphi_{t_0} = 0$ матрицы Φ (или может обнуляться строка $\theta_{t_0} = 0$ матрицы Θ), то тема t_0 исключается из тематической модели. Таким образом, в постановку задачи закладывается возможность уменьшать количество тем до оптимального, при заданных соответствующих регуляризаторах. Если для фиксированного документа d_0 обнуляется столбец $\theta_{d_0} = 0$ матрицы Θ, то документ d_0 фактически исключается из коллекции. Таким образом, регуляризованная модель может отказываться определять тематику документа d_0, если он не релевантен тематике коллекции или слишком короткий, при заданных соответствующих регуляризаторах.

В байесовских методах обучения тематических моделей [50, 52, 51] регуляризатор $R(\Phi, \Theta)$ интерпретируется как логарифм априорного распределения, а оптимизационная задача (8) соответствует принципу максимума апостериорной вероятности. В APTM регуляризатор не обязан иметь вероятностную интерпретацию.

3 Онлайновый EM-алгоритм

Для обработки больших коллекций лучше всего подходят онлайновые алгоритмы Online PLSA [45] и Online LDA [46]. Онлайновый EM-алгоритм реализует очень эффективную схему вычислений, при которой большие коллекции документов могут обрабатываться за небольшое количество итераций. Это происходит благодаря тому, что в матричном разложении $\Phi\Theta$ матрица Φ зависит от всей коллекции, тогда как в матрице Θ каждый столбец зависит от одного отдельного документа. Таким образом, онлайновый алгоритм основан на следующей стратегии обновлений. Итерации θ_{td} со встроенным E-шагом производятся при фиксированной матрице Φ для каждого документа d до сходимости. На последней итерации документа производится накопительное обновление переменных n_{wt}. Обновления матрицы Φ по переменным n_{wt} происходят по окончании обработки пачки (batch) документов. Если коллекция избыточна, что её тематика уже неплохо определяется по небольшой доле документов, и матрица Φ успевает сойтись задолго до того, как заканчивается первая итерация. Это позволяет эффективно применять онлайновые алгоритмы для анализа больших коллекций документов и новостных потоков.

Ниже представлен Online EM-алгоритм 3 для модели APTM, показывающий вычисления для коллекции документов D, разбитой на пачки (batch) документов D_b, $b = 1, \ldots, B$. Обработка каждого пакета выполняется алгоритмом 4.

Algorithm 3 Online EM-алгоритм для модели APTM

Вход: коллекция $D \equiv \{D_b\}_{b=1}^{B}$, коэффициент дисконтирования $\rho \in (0, 1]$;
Выход: конечная матрица распределения Φ;

инициализировать φ_{wt} для всех $w \in W$, $t \in T$;
обнулить $n_{wt}, \tilde{n}_{wt}$ для всех $w \in W$, $t \in T$;
повторять
  для всех пачек D_b, $b = 1, \ldots, B$:
    $(\tilde{n}_{wt}) := (\tilde{n}_{wt}) +$ *Обработать пакет документов* (D_b, Φ);
    если пора выполнить синхронизацию **то**
      $n_{wt} := \rho n_{wt} + \tilde{n}_{wt}$ для всех $w \in W$, $t \in T$;
      $\varphi_{wt} := \underset{w \in W}{\text{norm}}\left(n_{wt} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}}\right)$ для всех $w \in W$, $t \in T$;
      обнулить $\tilde{n}_{wt}$ для всех $w \in W$, $t \in T$;
пока Θ, Φ не сойдутся;

Эффективной параллельной реализацией Online EM-алгоритма для модели APTM является проект с открытым кодом BigARTM [42, 43] (`http://bigartm.org`), с которым мы в дальнейшем и будем работать. BigARTM разрабатывался исходя из требований асинхронной обработки данных, минимизации используемого объёма оперативной памяти, кроссплатформености и возможности работать как на одной машине, так и масштабируемости при увеличении количества ядер на узле.

Идея параллельного онлайнового алгоритма заключается в том, что каждый пакет документов D_b может обрабатываться параллельно при фиксированной матрице Φ, затем полученные на разных вычислителях обновления объединяются, и обновлённая версия матрицы Φ снова рассылается по вычислителям.

При этом одной из основных проблем работы с большими и потоковыми данными является адаптивный подбор вектора коэффициентов регуляризации $\vec{\tau}$.

Algorithm 4 Алгоритм обработки одного пакета документов

Вход: пачка документов D_b, матрица Φ;
Выход: матрица $(\tilde{n}_{wt})$;

обнулить $\tilde{n}_{wt}$ для всех $w \in W,\ t \in T$;
для всех документов $d \in D_b$:
  инициализировать $\theta_{td} := \frac{1}{|T|}$ для всех $t \in T$;
  повторять
    $p_{tdw} := \underset{t \in T}{\operatorname{norm}} (\varphi_{wt}\theta_{td})$ для всех $w \in d,\ t \in T$;
    $n_{td} := \sum_{w \in d} \tau_w n_{dw} p_{tdw}$ для всех $t \in T$;
    $\theta_{td} := \underset{t \in T}{\operatorname{norm}} \left(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right)$ для всех $t \in T$;
  пока θ_d не сойдётся;
  $\tilde{n}_{wt} := \tilde{n}_{wt} + \tau_w n_{dw} p_{tdw}$ для всех $w \in d,\ t \in T$;

Эффективно решать данную проблему и призвана текущая работа.

4 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 дальше — примеры регуляризаторов для выделения предметных и фоновых тем, а также регуляризатор увеличивающий различие тем.

О многих других регуляризаторах для АРТМ можно посмотреть в работах [44, 38, 40, 41].

4.1 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 теперь 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 это 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 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 *сглаживающим регуляризатором Дирихле*, т.к. эквивалентен предположению, что столбцы матриц Φ и Θ порождаются априорными распределениями Дирихле с гиперпараметрами $\beta_0\beta_w$ и $\alpha_0\alpha_t$ [1].

4.2 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 нашими распределениями φ_t и θ_d и равномерными $\beta = (\beta_w)_{w\in W}$, $\alpha = (\alpha_t)_{t\in T}$, 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* и он является противоположностью по оптимизационной задаче сглаживающего регуляризатора [1].

4.3 Декоррелирующий регуляризатор

Тематические модели интересны при различии тем. Понятие различности тем можно пытаться по-разному формализовать, например, как дискретных распределений $\varphi_{wt} = \mathsf{P}(w\,|\,t)$ или как нормированных векторов $\varphi_w = (\varphi_{wt})_{t\in T}$. Обычно в качестве понятия различности тем используют ковариацию:

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

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 [44, 47].

4.4 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 документов.

5 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, многократно изменяться в ходе исследования.

Критерии качества тематических моделей принято делить на внутренние (intrinsic) и внешние (extrinsic) [44]. 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 АРТМ можно посмотреть в работах [44, 38, 40, 41].

5.1 Перплексия

Перплексия — это мера несоответствия модели $\mathsf{P}(w\,|\,d)$ токенам w, она определяется через логарифм правдоподобия (2):

$$\mathscr{P}(D;p) = \exp\left(-\frac{1}{n}L(\Phi,\Theta)\right) = \exp\left(-\frac{1}{n}\sum_{d\in D}\sum_{w\in W} n_{dw}\ln \mathsf{P}(w\,|\,d)\right), \tag{10}$$

где $n = \sum\limits_{d\in D}\sum\limits_{w\in W} n_{dw}$ — 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.

5.2 Разреженность матриц

Как говорилось в подсекции 4.2, 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.

5.3 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);

$\text{kernel}_\text{t}\ \text{size} = |W_t|$ — размер ядра (ориентировочный оптимум $\frac{|W|}{|T|}$).

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

6 Постановка проблемы

Линейную комбинацию оптимизируемого функционала $L(\Phi,\Theta)$ и r регуляризаторов $R_i(\Phi,\Theta)$ с неотрицательными коэффициентами τ_i из задачи (8) можно переписать в виде:

$$R(\Phi) = \sum_{i=1}^{r} \vec{\tau}_i R_i(\Phi), \quad L(\Phi) + R(\Phi) \to \max_{\Phi}. \tag{11}$$

Точкой в данном функционале является только матрица Φ, на основе которой вычисляется матрица Θ при обработке пачки документов D_b используя *алгоритм 4 обработки пакета документов* (D_b, Φ).

Для решения данной оптимизационной задачи (11) 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.*

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 Φ^* функционала $L(\Phi) \to \max$, имеющая желаемые значения показателей качества $\vec{Q}^*$.

На каждой итерации обработки пачки документов l фиксируется вектор коэффициентов регуляризации $\vec{\tau}^l$, обрабатывается одна пачка $D_{b(l)}$, и по данным этой пачки и текущим приближениям матриц Φ и Θ вычисляется вектор показателей качества $\vec{Q}^l$.

Ниже приведён псевдокод алгоритма 5, который адаптивно управляет траекторией регуляризации.

Далее, текущие пакеты 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 5 теперь дополнительно делает разбиение коллекции на группы пачек $D \equiv \{D_p\}_{p=1}^P$, на которых будут происходить итерации метода, а каждая группа состоит за 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 3 для модели APTM*, в который мы можем включать более сложные механизмы регуляризации APTM моделей,

Algorithm 5 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: $\quad D_{b(l)} :=$ *взять следующую пачку из* $\{D_b\}_{b=1}^{B}$;
3: $\quad \Phi^{l+1} := F_l(\Phi^l, \vec{\tau}^l, D_{b(l)})$;
4: $\quad \vec{Q}^{l+1} := \vec{Q}(x^{l+1})$;
5: $\quad \vec{\tau}^{l+1} :=$ *выбрать следующий* $\vec{\tau}(\vec{Q}^{l+1}, \vec{\tau}^l, \Phi^{l+1}, D_{b(l)})$;

такие как включение или отключение некоторых регуляризаторов с определённой внутренней итерации Online EM-алгоритма.

Более того, если выборка документов очень разнообразна, то данный способ даёт возможность параллельно обрабатывать каждую итерацию l, тем самым позволяя брать большие группы пачек документов D_p, чтобы не получать слишком зашумлённые результаты из-за переобучения под конкретные документы, а выучивать более общую модель на каждой итерации.

Таким образом, данное объединение документов в группы пакетов позволяет получать более точные значения показателей качества $\vec{Q}$, улучшать сходимость APTM моделей и даёт гибкий способ настройки регуляризации APTM моделей.

Вдобавок, мы получаем алгоритм управления траекторией регуляризации модели APTM как надстройку над существующими Online EM-алгоритмами, что позволяет просто реализовать её поверх существующих решений и без труда использовать на практике.

Ниже приведён алгоритм 6, дополнительно группирующий пачки документов для лучшей сходимости.

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

Таким образом, нам остаётся разобраться с методом *выбора следующей точки вектора регуляризации* $\vec{\tau}$. При подборе пути регуляризации существуют следующие особенности.

Во-первых, показатели качества Q_j, определённые значения которых требуется достичь, могут лишь косвенно быть связаны с регуляризаторами R_i, которые участвуют в оптимизационной задаче непосредственно.

Во-вторых, исходная функция $a : \vec{\tau} \to \vec{Q}$ может является трудновычислимой, недифференцированной и даже негладкой из-за сложных регуляризаторов, что ограничивает методы решения данной оптимизационной задачи.

Далее, точка в пространстве показателей качества $\vec{Q}$ определяется неоднозначно, с учётом погрешности ε, которую даёт тематическая модель APTM, производящая

разложение на матрицы Φ и Θ при обработке конечной пачки документов $D_{b(l)}$, что накладывает определённые трудности при решении оптимизационной задачи.

Коэффициенты регуляризации $\vec{\tau}_i$ могут входить в противоречия и требовать изменений в разные стороны.

И последнее, при решении текущей оптимизационной задачи стоит учитывать, что вычисления происходят в пространствах небольших размерностей значений $\vec{Q}$ и $\vec{\tau}$.

В случаях больших и сильно избыточных коллекций для обучения тем и определения матрицы Φ может хватать одного прохода по коллекции, но в этом случае могут быть построены разные тематические модели для каждой группы пачек документов, в зависимости от порядка обработки её разбиения на пакеты документов. Таким образом, неоднородность модели по документам коллекции может плохо сказаться на качестве решения задач текстовой аналитики. Эта проблему можно решать путём обработки коллекции в два прохода. При первом проходе производятся исследовательские эксперименты и адаптируются коэффициенты регуляризации, определяется матрица тем Φ, а при втором проходе тематические модели всех документов строятся при фиксированной матрице Φ.

Часть III
Обзор выбора траектории регуляризации

Для управления траекторией регуляризации в рамках данной работы предлагается рассмотреть несколько методов *выбора следующей точки вектора регуляризации* $\vec{\tau}$, которые будут использоваться в алгоритме 5. Такие методы как, случайный поиск по фиксированной сетке и жадный поиск с экспертным выбором следующей точки сейчас используются на практике, при этом итерации алгоритма 5 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 не позволяет использовать данный метод на потоковых данных.

В текущих экспериментах делается попытка модифицировать случайный поиск и жадный поиск путём использования аппарата множественной регрессии и активного обучения, уходящего своими корнями в теорию оптимальных экспериментов [61, 62], тем самым позволяя автоматизировать процесс управления регуляризацией.

Помимо упомянутых выше методов будет рассматриваться случайный поиск и симплекс метод Нелдера—Мида [56] в качестве базовых решений, которые хорошо зарекомендовали себя при решении оптимизационных задач, где приходится ограничиваться только подсчётом значений функций и не использовать производные.

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 нашем 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.

7 Случайный поиск по сетке

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.

8 Симплекс метод Nelder—Mead

Симплекс метод Нелдера—Мида [56] — 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 деформи-

ровать 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).

9 Модификации жадного поиска

Экспертный выбор наилучшего следующего вектора коэффициентов регуляризации $\vec{\tau}$ путём анализа результатов предыдущих запусков сейчас является одним из основных методов на практике. Естественным способом автоматизировать этот процесс можно через множественную регрессию, т.е. восстановить множественную регрессионную зависимость вектора параметров $\vec{\tau}$, как зависимых переменных, от вектора показателей качества $\vec{Q}$, как независимыми переменными.

Различные стандартные методы решения задачи восстановления регрессии хорошо описаны в курсе лекций [57].

В ходе итерационного процесса накапливается обучающая выборка пар векторов $(\vec{Q}^l, \vec{\tau}^l)$, где вектор $\vec{Q}^l$ — это значения показателей качества, полученные при построении АРТМ модели с коэффициентами регуляризации $\vec{\tau}^l$. Если зависимость $\vec{\tau}(\vec{Q})$ удалось хорошо восстановить, то искомый вектор коэффициентов регуляризации $\vec{\tau}^*$ можно получить, вычислив как $\vec{\tau}^* := \vec{\tau}(\vec{Q}^*)$, где в качестве $\vec{Q}^*$ передаются желаемые значения показателей качества.

При восстановлении множественной регрессионной зависимости существуют следующие сложности.

Во-первых, изначально наша выборка $\{(\vec{Q}^l, \vec{\tau}^l)\}_{l=0}^N$ пуста и мы не можем сразу предсказать параметры для $\vec{Q}^*$. Основная задача состоит в том, чтобы как можно раньше направить итерационные 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 строиться по обучающей выборке, которая находится далеко от целевых значений $(\vec{Q}^*, \vec{\tau}^*)$. Таким образом, аппроксимация регрессионной моделью может давать недостаточную точность. В текущем алгоритме, предлагается не пытаться сразу предсказывать точку оптимума $\vec{Q}^*$, а делать шаги в её направлении $\vec{\tau}^{l+1} = \vec{\tau}((1-\alpha)\vec{Q}^l + \alpha\vec{Q}^*)$, где $\alpha \in [0, 1]$.

Более того, этого тоже может не хватать после нескольких итераций, если регрессионную зависимость в окрестности $\vec{Q}^*$ восстановить ещё не удалось, а на новом шаге l уже не выбирается существенно новая точка и не вносятся значимые изменения в регрессионную модель. Для решения этой проблемы предлагается чередовать *исследовательские* и *максимизирующие* шаги, чтобы исследовать ещё непокрытое регрессионной зависимостью прострнство 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. В секции 11 будет рассмотрен другой, более эффективный метод исследования, позаимствованный из активного обучения.

10 Модификации случайного поиска по сетке

Случайный поиск по сетке (см. секцию 7) тоже можно существенно улучшить с помощью множественной регрессии. В этом методе предлагается использовать зависимость $\vec{Q}(\vec{\tau})$, в предположении, что она может восстанавливаться лучше, чем зависимость $\vec{\tau}(\vec{Q})$. Таким образом, мы хотим восстанавливать регрессионную зависимость вектора показателей качества $\vec{Q}$ от вектора параметров $\vec{\tau}$.

В ходе итерационного процесса, аналогично секции 9, 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. 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 наша выборка $\{(\vec{\tau}^l, \vec{Q}^l)\}_{l=0}^N$ пуста и мы не можем сразу хорошо предсказать 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 строиться по обучающей выборке, 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. В секции 11 будет рассмотрен другой, более эффективный метод исследования, позаимствованный из активного обучения.

11 Модификации с активным обучением

Основной проблемой в обучении регрессионных моделей в секциях 9 и 10 является достижение окрестности целевого вектора значений показателей качества $\vec{Q}^*$. Чтобы модель могла аппроксимировать в этой окрестности с хорошей точностью, необходимо делать исследовательские шаги, которые направлены на улучшение способности к обобщению. При этом, мы хотим прийти в точку оптимума, сделав как можно меньше исследовательских шагов. Здесь и предлагается использовать идеи из активного обучения и теории оптимальных экспериментов [61, 62].

В данной работе предлагается использовать семплирование по несогласию в комитете [63]. В этом методе предлагается рассматривать несколько регрессионных моделей вместо одной и выбирать претендентом на исследование тот объект, на кото-

ром предсказания моделей расходятся сильнее всего. Расхождение между предсказанными векторами можно мерить, например, метрикой, которой мы планируем оценивать расстояние до целевого значения $\vec{Q}^*$ (см. подсекцию 13.3).

Часть IV
Эксперименты

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

Ниже описывается как происходило моделирование данных и приводятся некоторые рекомендации по настройке гиперпараметров и параметров системы.

12 Моделирование данных

Все эксперименты проводились на смоделированной коллекции документов. Выборка D состояла из 100000 документов, которые описывались словарём W из 2200 слов. Количество тем было взято $|T| = 10$, из которых было $|S| = 9$ 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 5.3), 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. Тематическую модель с фоновыми темами можно рассматривать как обобщение робастных моделей [54, 55], в которых использовалось только одно фоновое распределение.

В соответствии с этими предположениями были сгенерированы матрицы распределения слов в темах Φ и распределения тем в документах Θ.

Матрица Φ получилась разреженной для предметных тем с ядром около 150-250 слов для каждой темы, с минимальным пересечением тем по словам. В фоновую тему было взято около 1100 слов, которые составили все остальные общие слова, т.е. пересечения с предметными темами по этим словам не было.

Матрица Θ получилась разреженной по предметным темам, т.е. в каждом документе встречается не очень много предметных тем — в среднем по 1-2 темы на документ, со случайными пересечениями по темам между документами. Слова в фоновой теме распределены равномерно и встречаются в каждом документе.

Количество слов в документе было взято в среднем около 1000-1500.

13 Гиперпараметры системы

Правильная настройка гиперпараметров является очень важным начальным шагом, от которого зависят все дальнейшие результаты. Этот шаг требует вовлечения и экспертного понимания в тематических моделях, аддитивной регуляризации и показателях качества, которые планируются измерять.

13.1 Размер пакета документов и группы пакетов

Первым делом стоит разбить коллекцию документов D на пакеты D_b, где $b = 1, \dots, B$. Это разбиение требуется в параллельной версии онлайнового EM-алгоритма 3 для модели APTM.

Пакеты документов не должны быть слишком маленькими по размеру, т.к. каждая пачка рассылается на отдельную машину и участвует в обучении приближенной матрицы $\tilde{\Phi}$ алгоритмом 4. В дальнейшем эти пачки собираются со всех машин и участвуют в обновлении матрицы Φ. Таким образом, если в текущей пачке документов будет очень мало, то приближенная матрица $\tilde{\Phi}$ будет неинформативной, а издержки от пересылки пакетов будут значительными.

Пакеты документов также не должны быть излишне большими, т.к. тогда теряются преимущества онлайн алгоритма 3 при обработке больших коллекций.

В текущих экспериментах размер пакета (batch size) был выставлен в 200 документов, таким образом из 100000 документов было получено 500 пакетов.

Следующим параметром стоит задать количество пакетов в каждой группе (metabatch size). Полезно иметь в виду, что чем больше коллекция документов, тем больше её можно дробить, т.е. увеличивать число групп P. При этом, если каждый документ описывается большим числом слов, то тоже можно увеличивать P, т.к. сходимость может достигаться раньше. Но если необходимо выделить много тем T, то стоит наоборот уменьшать число групп P, т.к. необходимо больше итераций по документам внутри группы, чтобы все темы корректно описать.

В качестве эвристики, можно пользоваться следующим правилом:

$$P \propto \frac{|D| \frac{\sum\limits_{d_i \in D} |d_i|}{|D|}}{|T|} = \frac{\sum\limits_{d_i \in D} |d_i|}{|T|};$$

Для дополнительных экспериментов по качеству сходимости методов к целевому вектору показателей качества текущая коллекция документов разбивалась на 50 групп по 10 пакетов документов, на 20 групп по 25 пакетов документов и на 8 групп в среднем по 63 пакета документов.

13.2 Регуляризаторы и метрики качества

В текущих экспериментах будем использовать рассмотренные выше регуляризаторы (см. секцию 4) и метрики качества (см. секцию 5).

Подбор текущих регуляризаторов проводился на тех же принципах, что и генерирование модельных данных (см. секцию 12). Эти наблюдения полностью подтверждались при экспериментировании с комбинированием регуляризаторов сглаживания, разреживания и декоррелирования в APTM [44, 39, 40, 41] .

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

Текущий метод адаптивного подбора траектории регуляризации пытается построить предсказание с учителем, т.е. требует задать целевой вектор значений критериев качества $\vec{Q}^*$, который необходимо достичь. Это накладывает некоторые ограничения на применение данного метода, т.к. требуется заранее экспертно задавать целевые значения, а для некоторых показателей качества искомые значения могут быть неизвестны.

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

$$wMAE(\vec{Q}, \vec{Q}^*) = \frac{\sum_{j=0}^{r} w_j |\vec{Q}_j - \vec{Q}_j^*|}{m}; \tag{12}$$

Веса нужны для приведения значений показателей качества к единой шкале, 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 9, 10, 11 требуют дополнительных установок и уточнений по начальным и основным итерациям. Ниже описано какие конкретно использовались шаги и модели для исследований в текущих экспериментах.

14.1 Начальные итерации исследовательских методов

В методах секций 9, 10, 11 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 задавались экспертно, но если такой информации нет, то можно шагать по каждой координате в обе стороны.

Таким образом, мы получаем $m + 1$ пару векторов $(\vec{\tau}, \vec{Q})$, где m — это число используемых регуляризаторов для обучения АРТМ 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 секциях 9, 10 предлагалось брать случайную точку, а в секции 11 тот вектор, где больше всего расхождения комитета

регрессионных моделей. Все эти подходы будут использованы и сравнены в секции 15.

В качестве комитета будем рассматривать всего 2 модели восстановления регрессионной зависимости, — линейную регрессию *(Linear Regression)* и случайный лес *(Random Forest)* [58, 59, 60]. Линейная регрессия — это относительно простой алгоритм, который обладает большой обобщающей способностью. Случайный лес — это ансамбль решающих деревьев, который даёт хорошее качество в уже исследованной области.

Таким образом, на основных шагах исследования будет браться в зависимости от рассматриваемого метода либо случайная точка, либо точка, где прогнозируются самые непохожие результаты по метрике $wMAE$ (см. секцию 12) линейной регрессии и случайного леса, а на основных шагах максимизирования будет браться лучшее предсказание случайного леса.

15 Сравнение результатов

Ниже приведены результаты сходимости методов к целевому вектору значений показателей качества $\vec{Q}^*$ по метрике $wMAE$, рассмотренной в секции 12.

На рисунках изображена сходимость по метрике в зависимости от количества итераций по группам документов, где рассматриваются разные варианты разбиения модельной коллекции документов на группы: 50 групп по 10 пакетов документов, 20 групп по 25 пакетов документов и 8 групп по 63 пакета документов. Таким образом, проход по коллекции документов может осуществляться за разное число итераций, — 50, 20 или 8. А при фиксированных 50 итерациях методы будут проходить всю коллекцию от одного раза, в случае разбиения на 50 групп, до более чем 6 раз, в случае разбиения на 8 групп, что непосредственно является показателем затраченного времени. Тем не менее, графики от итераций по группам документов, а не от времени или числа обработанных документов, более наглядно показывают разницу в качестве сходимости методов при взятии разного числа документов на каждой итерации методов. Если это качество различается не существенно, то стоит отдавать предпочтение разбиениям на большее число групп для более быстрой сходимости.

На рисунках 1, 2, 3 представлены результаты базовых алгоритмов — случайный поиск (см. секцию 7) и симплекс метод Nelder—Mead (см. секцию 8), они дают базовые ориентиры на понимание принимаемых значений метрики $wMAE$ и оценки скорости сходимости методов.

На рисунках 4, 5, 6 рассмотрены графики сходимостей модифицированного жадного алгоритма (см. секцию 9), модифицированного случайного поиска по сетке (см. секцию 10) и их улучшений с использованием семплирования по несогласию в комитете (см. секцию 11). 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.

На рисунках 7, 8, 9 приведены графики сходимости максимизирующих шагов модифицированных методов в сравнении с оптимизационным методом Nelder—Mead.

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 регуля-

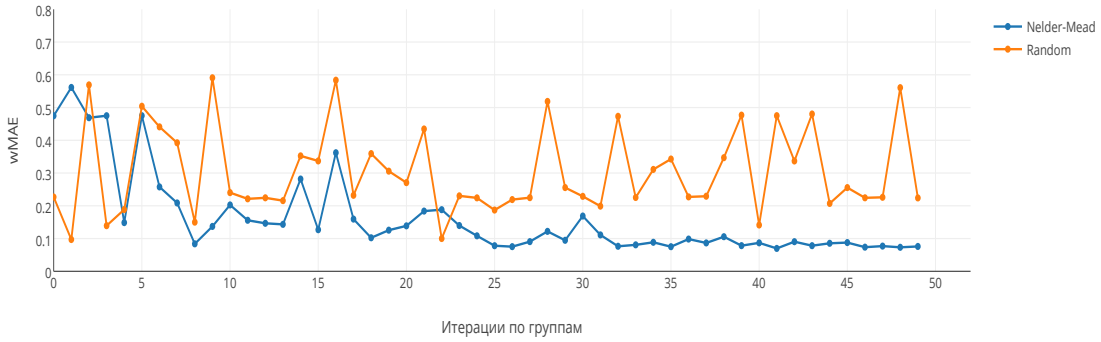

Рис. 1: График сходимости при разбиении на 8 групп

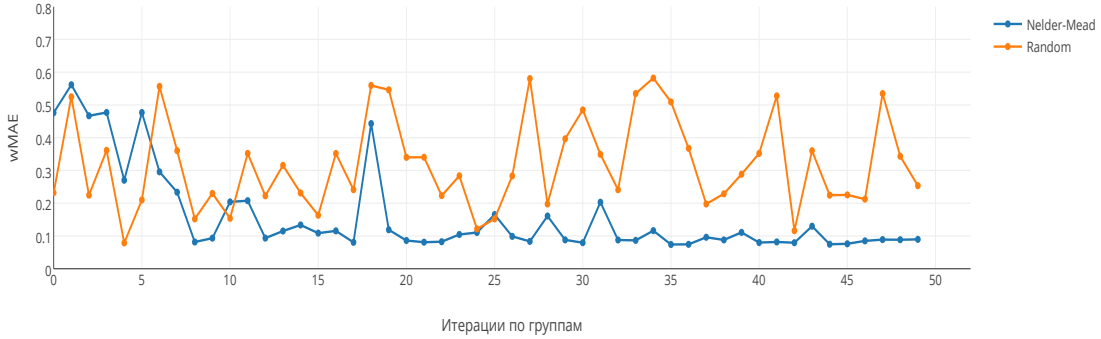

Рис. 2: График сходимости при разбиении на 20 групп

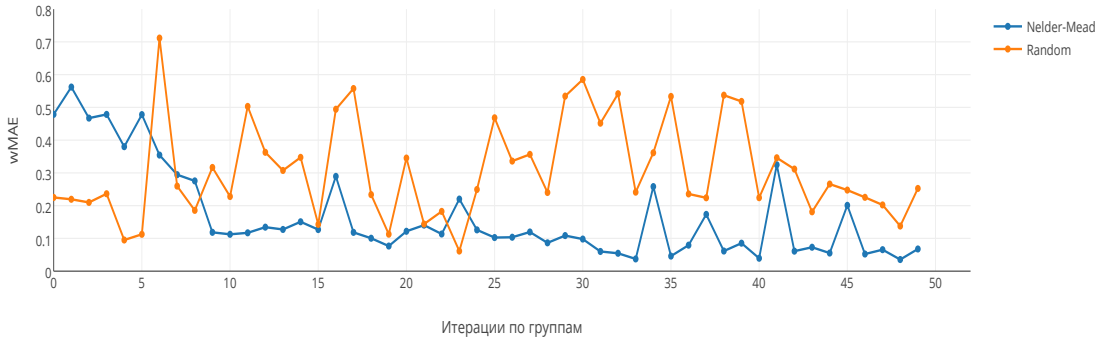

Рис. 3: График сходимости при разбиении на 50 групп

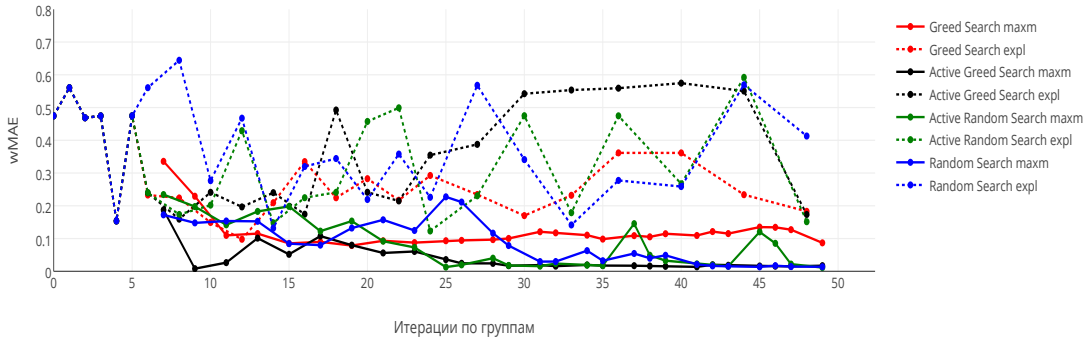

Рис. 4: График сходимости при разбиении на 8 групп

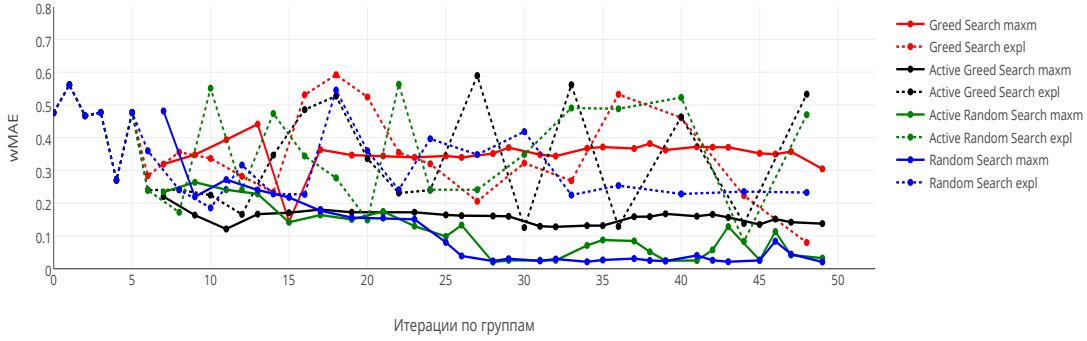

Рис. 5: График сходимости при разбиении на 20 групп

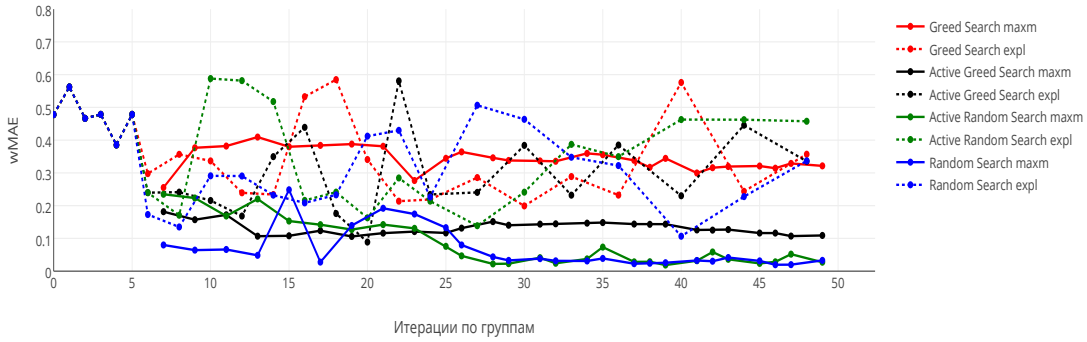

Рис. 6: График сходимости при разбиении на 50 групп

ризации $\vec{\tau}^l$.

Из рисунков видно, что модифицированные методы случайного поиска по сетке на основе восстановления регрессионной зависимости $\vec{Q}(\vec{\tau})$ показывают более хорошее качество по сравнению с симплекс методом и модифицированными методами жадного поиска, что подтверждает гипотезу о том, что прямая регрессионная зависимость $\vec{Q}(\vec{\tau})$ восстанавливается лучше, чем обратная $\vec{\tau}(\vec{Q})$.

Можно заметить, что на группах больших размеров методы ведут себя более стабильно и лучше сходятся, но на малых группах также наблюдается хорошая сходимость некоторых методов, хоть и более стохастическая, делающая при этом в несколько раз меньше обходов коллекции документов.

Более того, на исследовательских шагах вместо случайных точек лучше использовать несогласие в комитете, что даёт большую обобщающую способность регрессионным моделям и способствует лучшему восстановлению зависимости.

Стоит отметить, что алгоритм 5 адаптивного управления траекторией регуляризации не существенно увеличивает время обработки коллекции документов по сравнению с применением Online EM-алгоритма 3 для модели APTM. Наибольшее время в текущей надстройке уходит на выбор следующего вектора коэффициентов регуляризации $\vec{\tau}$, в частности на обучение и применение таких моделей по восстановлению регрессионной зависимости, как случайный лес. Как можно увидеть на рисунках 10, 11, методы с использованием случайного леса из 50 деревьев на максимизирующих шагах модифицированных методов не уступают по качеству сходимости методам со случайным лесом из 2000 деревьев, поэтому можно ограничиваться более простыми моделями, чтобы адаптивная траектория регуляризации не существенно замедляла Online EM-алгоритм.

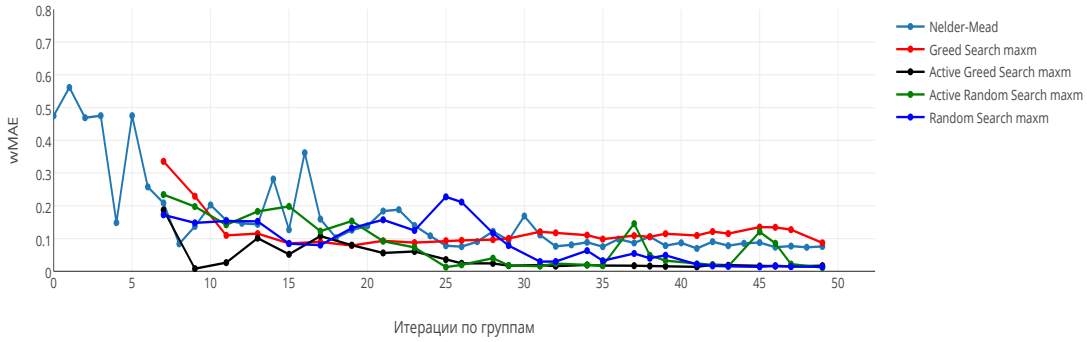

Рис. 7: График сходимости при разбиении на 8 групп

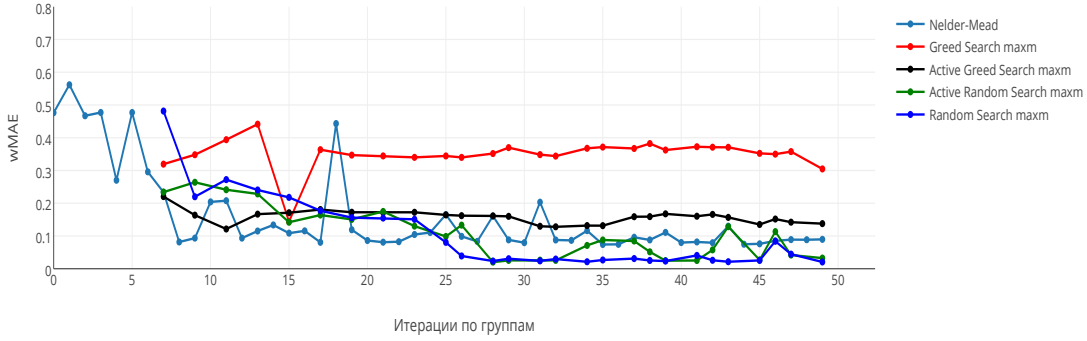

Рис. 8: График сходимости при разбиении на 20 групп

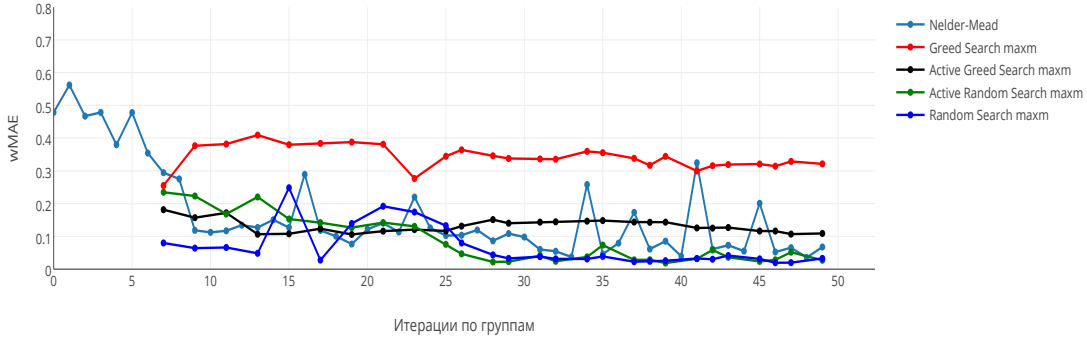

Рис. 9: График сходимости при разбиении на 50 групп

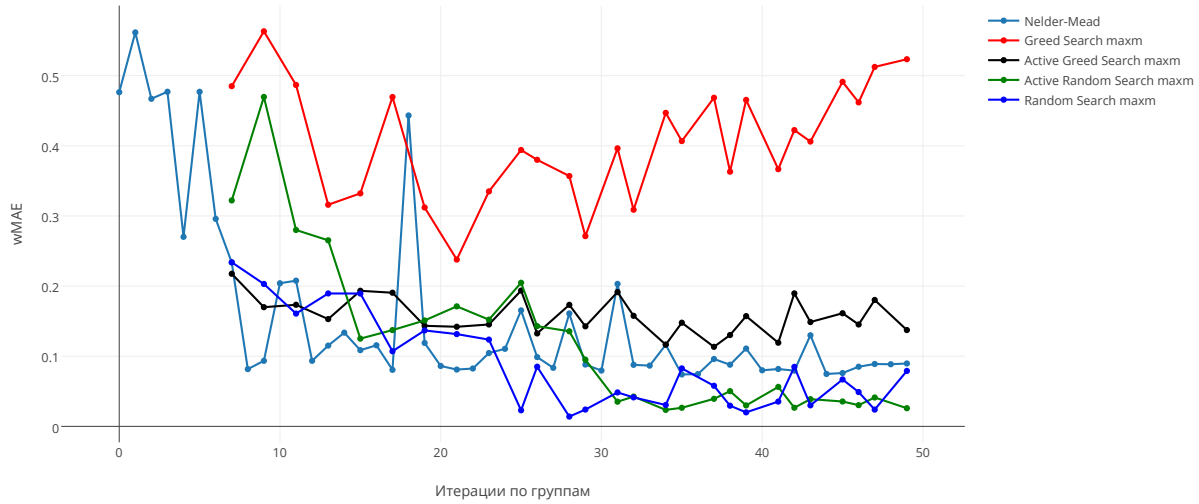

Рис. 10: График сходимости при разбиении на 20 групп (с обучением 50 деревьев)

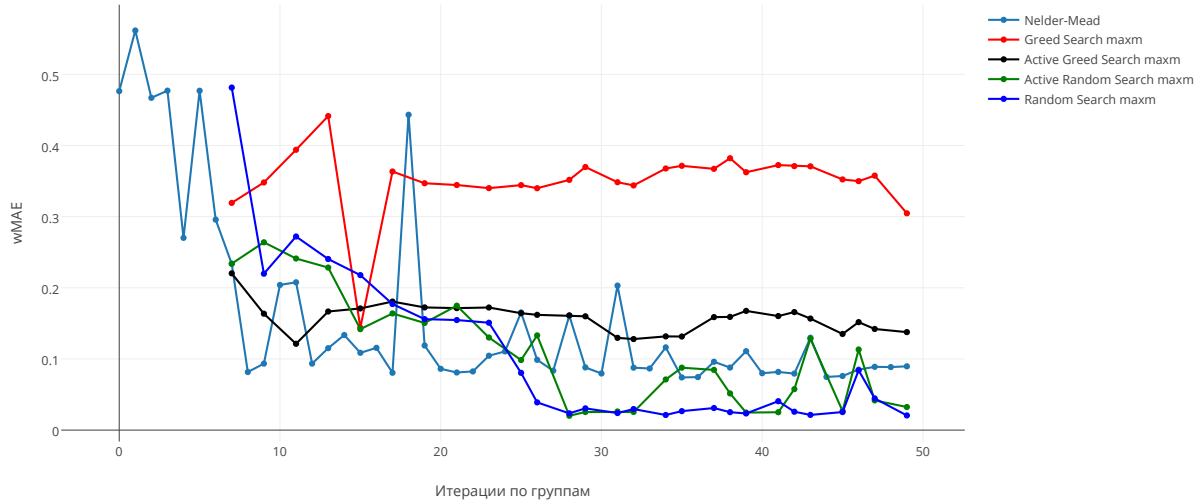

Рис. 11: График сходимости при разбиении на 20 групп (с обучением 2000 деревьев)

Часть V
Выводы

В данной работе разработаны новые эффективные методы адаптивного подбора траектории регуляризации в тематическом моделировании (см. часть III), которые естественным образом обобщают текущие экспертные методы подбора коэффициентов регуляризации.

Проведены эксперименты по сравнению различных методов выбора траектории регуляризации на модельных данных (см. часть IV). Стоит отметить, что самыми быстрыми сходимостями при решении текущей задачи обладают модификации случайного поиска по сетке (см. секцию 10).

Предлагаемое решение обладает как положительными свойствами, так и отрицательными.

К положительным свойствам можно отнести автоматизацию подбора траектории регуляризации и значительное ускорение сходимости методов, в том числе благодаря обучению методов на небольших частях коллекции документов. Эти свойства также позволяют применять тематическое моделирование на потоковых данных. Разработанные методы допускают подбор коэффициентов для произвольных регуляризаторов и могут настраиваться на различные функционалы качества, которые могут не обладать свойствами дифференцируемости и гладкости. Более того, данное решение обладает гибкой системой настройки коэффициентов регуляризации, т.е. легко можно изменять количество регуляризаторов при итерировании Online EM-алгоритмом модели APTM.

Предложенный метод обладает и недостатками. К отрицательным свойствам текущего подхода можно отнести то, что для эксплуатации требуются экспертные знания. Эксперту нужно задавать целевые значения показателей качества, которые необходимо достичь, и допустимые отклонения от этого качества в виде метрики. Кроме того, для модифицированных методов случайного поиска эксперту необходимо объявлять сетку возможных принимаемых значений коэффициентов регуляризации, от которой существенно будет зависить сходимость методов.

Кроме того, рассмотренный метод имеет параметры и гиперпараметры, от которых зависит качество и скорость сходимости. В данной работе предложены способы настройки гиперпараметров системы и параметров методов (см. часть IV).

Список литературы

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

[2] 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.

[3] 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.

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

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

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

[7] 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.

[8] 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.

[9] 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.

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

[11] 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.

[12] 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.

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

[14] 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.

[15] 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.

[16] 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.

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

[18] 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.

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

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

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

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

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

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

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

[26] 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.

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

[28] 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

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

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

[35] 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.

[36] 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.

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

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

[39] 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.

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

[41] 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.

[42] 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.

[43] 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).

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

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

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

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

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

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

[50] 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.

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

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

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

[54] Chemudugunta C., Smyth P., Steyvers M. Modeling general and specific aspects of documents with a probabilistic topic model // Advances in Neural Information Processing Systems. — MIT Press, 2007.

[55] Potapenko A. A., Vorontsov K. V. Robust PLSA performs better than LDA // 35th European Conference on Information Retrieval, ECIR-2013, Moscow, Russia, 24-27 March 2013. — Lecture Notes in Computer Science (LNCS), Springer Verlag-Germany, 2013.

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

[57] Воронцов К. В. Курс лекций по алгоритмам восстановления регрессии, — 2007.

[58] Ho T. K. Random Decision Forests. Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August 1995.

[59] Ho T. K. The Random Subspace Method for Constructing Decision Forests. IEEE Transactions on Pattern Analysis and Machine Intelligence, — 1998.

[60] Breiman L. Random Forests. Statistics Department. University of California. Berkeley, CA 94720. January 2001.

[61] Settles B. Active Learning Literature Survey, — 2010.

[62] Federov V. Theory of Optimal Experiments. Academic Press, — 1972.

[63] Seung H. S., Opper M., Sompolinsky H. Query by Committee. Racah Institute of Physics and. Center for Neural Computation. Hebrew University. Jerusalem 91904, Israel, — 1992.