Министерство образования и науки Российской Федерации
Московский физико-технический институт (государственный университет)
Факультет управления и прикладной математики
Кафедра «Интеллектуальные системы»

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА МАГИСТРА

Иерархическая мультимодальная тематическая модель коллекции научно-популярных текстов

Выполнила:
студентка 6 курса 174 группы
Ефимова Ирина Валерьевна

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

Москва, 2017

Содержание

Введение **3**

1 Постановка задачи **5**

1.1 Плоская модель 6

1.2 Иерархическая модель hARTM 7

1.2.1 Иерархический регуляризатор разреживания 8

1.3 Постановка задачи 9

2 Иерархическая модель коллекции научно-популярных текстов **10**

2.1 Задача построения первого уровня иерархии 11

2.2 Задача построения второго уровня иерархии 12

2.3 Задача автоматического именования тем 13

2.3.1 Признаки 14

2.3.2 Экспертная оценка именования тем 15

2.3.3 Оценка качества темы и ее названия 16

3 Вычислительный эксперимент **16**

3.1 Описание данных 16

3.2 Метрики качества моделей 16

3.2.1 Перплексия 17

3.2.2 Разреженность 17

3.2.3 Ошибка первого рода 17

3.2.4 Ошибка второго рода 18

3.3 Сравнение моделей 18

3.3.1 Построение первого уровня иерархии 18

3.3.2 Построение второго уровня иерархии 22

3.4 Автоматическое именование тем 23

3.5 Выводы и рекомендации 24

Заключение **26**

Аннотация

Рассматривается задача построения двухуровневой тематической иерархии с автоматическим именованием тем. Предполагается, что документы тегированы, то есть каждому документу редакторами ресурса приписано некоторое количество ключевых слов или фраз. Также при решении педполагается, что среди тегов находятся названия тем разных уровней иерархии мультимодальной тематической модели коллекции документов. В работе предлагается метод послойного построения иерархии текстовых коллекций и алгоритм автоматического именования тем.

Ключевые слова: *вероятностное тематическое моделирование; аддитивная регуляризация тематических моделей; ARTM; BigARTM; иерархическая тематическая модель; научно-популярные тексты*

Введение

В эпоху информационных технологий появляется доступ к неограниченному объему знаний, доступных через сеть Интернет, но физически человек способен ознакомиться только с малой частью этих данных. Возникает потребность в системах автоматической организации информации для пользователя.

В последнее время активно развивается раздел машинного обучения, решающий задачу поиска тем в коллекции документов — вероятностное тематическое моделирование. Тематическая модель коллекции текстовых документов определяет, к каким темам относится каждый документ и какие слова (токены) образуют каждую тему. Иными словами, модель задает компактное представление для коллекции, которое позволяет быстрее ознакомиться с ее содержанием. Однако на больших текстовых коллекциях, когда число тем становится равным нескольким сотням или тысячам, даже такое представление в виде набора тем перестает быть удобным. Появляется потребность в построении иерархических тематических моделей, в которых крупные темы постепенно дробятся на более узкие, специализированные темы. Таким образом, иерархические тематические модели помогают представить структуру коллекции текстовых документов в виде иерархии тем. Это позволяет пользователю наиболее полно познакомиться с областью знаний, к которой относится коллекция.

Большинство подходов к построению иерархий вероятностные: в них термины, темы и документы считаются случайными величинами, а коллекция моделируется с помощью процесса порождения слова в документе. Одна из первых таких иерархических моделей предложена в [1]: иерархия представляется в виде дерева тем, и ее можно достраивать при добавлении новых документов в коллекцию. В [6] ключевая идея состоит в отказе от ограничения на граф: иерархия является многодольным графом, то есть темы могут иметь несколько надтем. Авторы [16] также представляют иерархию в виде многодольного графа и описывают две модели, которые автоматически определяют количество тем и количество уровней, или долей в графе. Одна модель строит иерархию документов, другая – иерархию терминов, совместить эти модели в одной не предлагается. Аналогично, в [12] темы описываются только лексикой, то есть связь документов и тем не моделируется; ключевая особенность – темы представляются как список фраз, а не отдельных терминов, в результате по-

вышается интерпретируемость тем. Список терминов родительской темы получается объединением списков терминов дочерних тем. Этот подход развивается в [11], где модель учитывает не только текстовую, но и иную информацию, представленную в коллекции: авторов, метки времени, локации на карте и т. д. В [13] делают акцент на трех приоритетах: масштабируемость, то есть быстрое построение модели на больших коллекциях, устойчивость, то есть построение похожих моделей при повторных запусках, и интерпретируемость. В [14] к этому списку добавляется еще одна цель: возможность учитывать указания эксперта, например указание объединить две темы. На важность масштабируемости алгоритма обучения также указывают авторы [15].

Цель работы заключается в разработке методики автоматического построения хорошо интерпретируемых тематических иерархий научно-популярных текстов. При решении предполагается, что есть информация о тегах каждого документа коллекции, то есть некоторое количество ключевых слов или фраз. При этом для первого уровня иерархии известно количество тем и их названия, которые являются подмножеством множества тегов. Таким образом, задача построения тематической модели первого уровня иерархии равносильна классификации документов по заданным темам. Для оценивания качества классификации предложены метрики качества: ошибки первого и второго рода. При построении остальных уровней иерархий возникает задача автоматического именования тем. Для ее решения предлагается алгоритм, который основан на предположении, что в тегах документах содержатся названия тем всех уровней.

При построения иерархических тематических моделей используется метод послойного построения иерархии, описанный в [2], который основан на аддитивной регуляризации тематических моделей (Additive Regularization of Topic Models, ARTM) [9].

Для проведения экспериментов используется BigARTM — библиотека для тематического моделирования с открытым исходным кодом [4, 7].

1 Постановка задачи

Пусть D – множество документов. Каждый документ $d \in D$ может состоять из элементов различных модальностей: слов (униграмм), словосочетаний (биграмм, триграмм и т.д.), теов, меток времени, авторов и т.д. Каждой модальности соответствует отдельный словарь, состоящий из всевозможных значений элементов модальности. Множество модальностей будем обозначать M, а их словари – $W^m, m \in M$; $W = \bigcup\limits_{m \in M} W^m$. Каждый документ $d \in D$ представляет собой последовательность n_d терминов $(w_1, ..., w_{n_d})$, принадлежащих словарю W.

Предполагается, что существует конечное множество тем T, и каждое употребление термина w в каждом документе d связано с некоторой темой $t \in T$, которая неизвестна. В вероятностном тематическом моделировании коллекция документов рассматривается как множество троек (d, w, t), заданного на конечном множестве $D \times W \times T$. Документы $d \in D$ и термины $w \in W$ являются наблюдаемыми переменными, тема t является латентной (скрытой) переменной.

Также предполагается, что порядок терминов в тексте не важен для определения его тематики. В этом случае коллекцию можно представить в виде матрицы частот слов с элементами n_{dw} – частота вхождения термина w в документ d. По матрице частот слов можно оценить вероятности появления терминов в документе: $p(w|d) = \frac{n_{dw}}{n_d}$.

Задача построения тематической модели коллекции документов заключается в поиске распределения $p(w|t)$ для всех тем $t \in T$ и распределения $p(t|d)$ для всех документов $d \in D$.

Все методы решения поставленной задачи основаны на предположении, что появление слов в документе d, относящихся к теме t, описываются общим для всей коллекции распределением $p(w|t)$ и не зависит от документа d:

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

Это предположение называется *гипотезой условной независимости.*

1.1 Плоская модель

При сделанных предположениях плоская (одноуровневая) тематическая модель описывается формулой

$$p(w|d) = \sum_{t \in T} p(t|d)p(w|t), \;\; d \in D, \; w \in W^m,$$

которая следует из определения условной вероятности, формулы полной вероятности и гипотезы условной независимости.

Пусть

- $F^m = \{p(w|d)\}_{W^m \times D}, m \in M$ — матрицы наблюдаемых вероятностей для каждой модальности;
- $\Phi^m = \{\phi_{wt}\}_{W^m \times T}, m \in M, \; \phi_{wt} = p(w|t)$ — матрицы терминов тем;
- $F = \bigcup\limits_{m \in M} F^m$, $\Phi = \bigcup\limits_{m \in M} \Phi^m$;
- $\Theta = \{\theta_{td}\}_{T \times D}, \; \theta_{td} = p(t|d)$ — матрица тем документов.

Обычно число тем $|T|$ много меньше числа документов $|D|$ и числа терминов $|W|$, поэтому задача сводится к поиску приближённого представления заданной матрицы частот F в виде низкорангового матричного разложения

$$F \approx \Phi\Theta,$$

где параметрами модели являются матрицы Φ и Θ, столбцы которых представляют дескретные распределения вероятности. Такие матрицы называются стохастическими.

В ARTM [9] параметры модели предлагается настраиваются путём максимизации взвешенной суммы логарифмов правдоподобия и регуляризаторов с помощью EM-алгоритма:

$$\sum_{m \in M} \eta_m \sum_{d \in D} \sum_{w \in W^m} n_{dw} \ln \sum_{t \in T} \phi_{wt}\theta_{td} + \sum_i \tau_i R_i(\Phi, \Theta) \to \max_{\Phi, \Theta} \tag{1}$$

$$\sum_{w \in W^m} \phi_{wt} = 1, \; \phi_{wt} \geq 0; \;\; \sum_t \theta_{td} = 1, \; \theta_{td} \geq 0, \tag{2}$$

где регуляризаторы $R_i(\Phi, \Theta)$ – выражают дополнительные требования к модели (например, разреженность и различность тем, наличие фоновых тем общей лексики языка [9]), коэффициенты τ_i и η_m введены для балансирования важности критериев.

В [9] доказана следующая теорема:

Теорема 1 *Если $R(\Phi, \Theta)$ непрерывно дифференцируема, то стационарная точка задачи (1)-(2) удовлетворяет следующей системе уравнений:*

E-шаг: $p(t|d,w) = \underset{t\in T}{\text{norm}}(\phi_{wt}\theta_{td})$;

M-шаг: $n_{wt} = \sum\limits_{d\in D} \eta_{m(w)} n_{dw} p(t|d,w)$, $n_{td} = \sum\limits_{w\in W} \eta_{m(w)} p(t|d,w)$;

$$\phi_{wt} = \underset{w\in W^m}{\text{norm}}\left[n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right], \quad \theta_{td} = \underset{t\in T}{\text{norm}}\left[n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right];$$

где оператор norm *преобразует вещественный вектор в дискретное распределение:* $\underset{t\in T}{\text{norm}}(x_t) = \frac{\max\{x_t,0\}}{\sum\limits_{t\in T}\max\{x_t,0\}}$.

Применение метода простой итерации к данной системе уравнений дает EM-алгоритм для обучения модели: E- и M-шаги алгоритма чередуются до стабилизации логарифма правдоподобия. Параметры модели инициализируются случайно.

1.2 Иерархическая модель hARTM

Для построения иерархических тематических моделей в [2] в модель ARTM предлагается ввести специальный межуровневый регуляризатор. Предполагается, что каждый уровень иерархии представляет собой плоскую модель ARTM.

Предположим, что построено $l \geq 1$ уровней тематической иерархии. Параметры модели будем обозначать следующим образом: Φ_l – матрица терминов тем, Θ_l – матрица тем документов, T_l – множество тем l-го уровня. Тогда связь между родительским и дочерним уровнями можно записать в виде:

$$\Phi_l \approx \Phi_{l+1}\Psi,$$

где $\Phi_{l+1} = \{p(w|t_{l+1})\}_{W\times T_{l+1}}, \Psi = \{p(t_{l+1}|t_l)\}_{T_{l+1}\times T_l}$, T_{l+1} – множесто тем $(l+1)$-го уровня (дочернего уровня). Матрица перехода Ψ содержит распределения тем $t_{l+1} \in T_{l+1}$ $(l+1)$-го уровня в темах $t_l \in T_l$ l-го уровня, то есть она показывает вероятность перехода из темы t_l в подтему t_{l+1}.

Если в качестве меры близости распределений использовать дивергенцию

Кульбака-Лейблера, то регуляризатор примет вид ([2]):

$$R(\Phi_{l+1}, \Psi) = \sum_{t_l \in T_l} \sum_{w \in W} n_{wt_l} \ln \sum_{t_{l+1} \in T_{l+1}} \phi_{wt_{l+1}} \psi_{t_{l+1}t_l} \to \max_{\Phi, \Theta}. \tag{3}$$

Формула описанного регуляризатора имеет схожую структуру с формулой правдоподобия модели. Поэтому такая постановка задачи эквивалентна добавлению в коллекцию $|T_l|$ псевдодокументов, представленных матрицей $\{n_{wt_l}\}_{W \times T_l}$. Тогда Ψ образует $|T_l|$ дополнительных столбцов матрицы Θ, соответствующих этим псевдодокументам.

Вес данного регуляризатора будем обозначать серез $\varkappa_1$.

1.2.1 Иерархический регуляризатор разреживания

В предложенной выше иерархической модели предполагается, что каждая тема с уровня $(l+1)$ может иметь несколько родительских тем с уровня l. По иерархической модели легче осуществлять поиск, если число тем $(l+1)$-го уровня невелико. Другими словами, ожидается, что распределение $\tilde{\boldsymbol{\psi}_{t_{l+1}}} = \{p(t_l|t_{l+1})\}_{t_l \in T_l}$ для всех $t_{l+1} \in T_{l+1}$ будет разреженным. Для этого в работе [2] предлагается иерархический регуляризатор разреживания, который оказывает влияние на элементы матрицы Ψ. Действительно, $p(t_l|t_{l+1})$ выражается через $p(t_{l+1}|t_l)$ (элементы матрицы Ψ) по формуле Байеса:

$$p(t_l|t_{l+1}) = \frac{p(t_{l+1}|t_l)p(t_l)}{\sum_{t'_{l+1}} p(t'_{l+1}|t_l)p(t_l)} = \frac{\psi_{t_{l+1}t_l}p(t_l)}{\sum_{t'_{l+1}} \psi_{t_{l+1}t_l}p(t_l)}.$$

Для разреживания распределения $\tilde{\boldsymbol{\psi}}_{t_{l+1}}$ в [2] предлагается максимизировать расстояние Кульбака-Лейблера между данным распределением и равномерным $\boldsymbol{\gamma} = \{\frac{1}{|T_l|}\}_{t_l \in T_l}$:

$$\sum_{t_{l+1} \in T_{l+1}} KL(\boldsymbol{\gamma}, \tilde{\boldsymbol{\psi}}_{t_{l+1}}) \to \max_{\Psi}.$$

Тогда регуляризатор будет выглядеть следующим образом:

$$R(\Psi) = \sum_{t_{l+1} \in T_{l+1}} \sum_{t_l \in T_l} \frac{1}{|T_l|} \ln p(t_l|t_{l+1}) = \frac{1}{|T_l|} \sum_{t_l} \sum_{t_{l+1}} \frac{\psi_{t_{l+1}t_l}p(t_l)}{\sum\limits_{t'_l} \psi_{t_{l+1}t'_l}p(t'_l)} \to \min_{\Psi}, \tag{4}$$

где $p(t_l)$ вычисляется по Θ^l.

В результате работы данного регуляризатора разреживания (4) вероятности $p(t_l|t_{l+1})$ для темы $t_{l+1} \in T_{l+1}$, имеющие большие значения для $t_l \in T_l$, становятся

еще больше в ходе итераций, а имеющие меньшие значения для $t_l \in T_l$, становятся еще меньше.

Также данный регуляризатор позволяет исключить случаи, когда тема с $(l+1)$-го уровня стремится стать дочерней для всех тем с l-го уровня:

$$p(t_{l+1}|t_l) \to \frac{1}{|T_l|}, \ \ \forall t_l \in T_l.$$

Теоретически возможна ситуация, когда модели не удается для темы $t_{l+1} \in T_{l+1}$ выделить хотя бы одну тему $t_l \in T_l$: $p(t_{l+1}|t_l) > 0$. В этом случае возникает "тема-сирота", у которой нет ни одной родительской. Тогда надо уменьшить коэффициент регуяризации для разреживающего регуляризатора (4). В наших экспериментах такие ситуации не возникали.

Вес данного регуляризатора будем обозначать через $\varkappa_2$.

1.3 Постановка задачи

Иерархические рубрикаторы позволяют пользователям быстрее находить нужные тематические разделы и новые документы, релевантные интересам позльзователя. В настоящее вермя сущестует много иерархических рубрикаторов (библиотечные классификаторы УДК, ББК, ГРНТИ, классификаторы фондов РФФИ, РНФ). Однако каждый из них имеет свои цели и задачи, многие рубрикаторы слишком громоздкие или фрагментарно устаревшие для применения к современному контенту в Интернете. Поэтому цель данной работы состоит в разработке методики автоматического построения хорошо интерпретируемых тематических иерархий для коллекции научно-популярных текстов.

В рассматриваемой задаче предполагается, что документы тегированы, то есть каждому документу редакторами ресурса приписано некоторое количество ключевых слов или фраз. Предполагается, что среди тегов находятся названия тем разных уровней иерархии.

Множество тегов вынесем в отдельную модальность. Пусть G — словарь данной модальности, $g \in G$ — её элементы (теги).

Рассмотрим задачу построения двухуровневой тематической иерархии с автоматическим именованием тем. При решении данной задачи возникают три подзадачи.

1. Задача построения первого уровня иерархии. Предполагается, что число тем $|T|$ верхнего уровня и их названия заданы экспертами — редакторами научного ресурса. Требуется построить тематическую модель, решив систему (1)-(2), где среди модальностей есть модальность тегов G.

2. Задача построения второго уровня иерархии. Требуется решить систему (1)-(2), где среди модальностей есть модальность тегов G, а среди регуляризаторов есть межуровневый (3), связывающий темы второго уровня с темами верхнего уровня.

3. Задача автоматического именования тем. Необходимо каждой теме $t \in T$ поставить в соответствие название из множества тегов G, наилучшим образом описывающее данную тему t.

2 Иерархическая модель коллекции научно-популярных текстов

Пусть Φ^g – подматрица матрицы Φ, соответствующая модальности тегов.

Пусть S – множество предметных тем, B – множество фоновых тем, $T = S \cup B$. Предполагается, что $S \subset G$. В общем случае каждая тема может быть представлена несколькими тегами. D_t — документы коллекции D, содержащие тему $t \in T$, G_d — множество тегов документа d.

Определим фоновость документа:

$$b(d) = \sum_{t \in B} \theta_{td}. \tag{5}$$

Фоновость документа определяет долю слов документа, относящихся к словам общей лексики, то есть не характеризующих никакие предметные темы.

Предлагается на каждом уровне иерархии вводить фоновые темы, так как это способствут очищению предметных тем от общей лексики языка и повышает качество модели [8, 10].

Темы l-го уровня будем обозначать $T_l = S_l \cup B_l$.

2.1 Задача построения первого уровня иерархии

Для классификации документов $d \in D$ по заданным предметным темам $S_1 \subset T_1$ предлагается сделать следующие шаги:

1. Обычно число тегов $|G|$ много меньше числа слов. Поэтому модальность тегов в уравнении (1) предлагается учитывать с большим весом η_g ($\approx 10^2$, при условии, что вес модальности слов равен 1). Иначе модальность тегов не будет вносить вклад в модель.

2. Предлагается матрицу Φ^g инициализировать специфическим образом (инициализация матриц Φ и Θ проводится до построения модели и по умолчанию Φ инициализируется случайными значениями, а Θ — константными $\theta_{td} = p(t|d) = \frac{1}{|T|}$, при этом $\sum\limits_{w \in W^m} \phi_{wt} = 1$ для всех $m \in M$, $\sum\limits_{t \in T} \theta_{td} = 1$):

$$\phi_{gt} = \begin{cases} z, z > 0, & \text{если } g = t, g \in G_1 \\ 0, & \text{если } g \neq t, g \in G_1 \\ \text{rand}(0,1), & \text{если } g \in G \setminus G_1 \end{cases} \tag{6}$$

   где $G_1 \subset G$ образуют теги, которые выбраны в качестве названия тем первого уровня, то есть $|G_1| = |S_1|$, а затем сделать перенормировку таким образом, что $\sum\limits_{g \in G} \phi_{gt} = 1 \;\; \forall t \in S_1$

Если при инициализации элементам матриц Φ и Θ присвоить нулевые значения, то и в конечной модели эти элементы останутся нулевыми. Благодаря такому свойству предложенная инициализация позволяет строить разнородные темы без пересечений.

Ненулевые значения z, при условии выполнения первого шага, позволяют к заданным темам S_1 притягивать токены всех модальностей данной тематики. На первый взгляд кажется вполне логичным в формуле (6) использовать $z = 1$, но в этом случае модель будет вырожденной и не будет возможности оставшимся тегам $g \in G \setminus G_1$ распределиться по темам первого уровня S_1, что очень важно для построения следующих уровней иерархии $l > 1$.

2.2 Задача построения второго уровня иерархии

Гипотеза 1. Количество фоновых тем второго уровня больше по сравнению с предыдущим уровнем иерархии: $|B_2| > |B_1|$. При этом:

- среди тем B_2 есть $|B_1|$ тем, которые наследуют темы B_1 — B_{21}. То есть каждой теме $t_1 \in B_1$ ставится в соответствие тема $t_2 \in B_{21}$, такая, что t_2 является единственной дочерней темой t_1, а t_1 является единственным родителем t_2;
- оставшиеся темы из B_2 являются фоновыми темами плоской модели, описывающей второй уровень иерархии.

Действительно, если данную гипотезу не принимать во внимание, то может возникнуть ситуация, когда фоновая тема первого уровня является одним из родителей (а порой и единственным) нескольких предметных тем второго уровня.

Для формализации данной гипотезы предлагается ввести специальные регуляризаторы разреживания матрицы Ψ, являющейся подматрицей матрицы Θ, таким образом, чтобы в построенной модели выполнялось:

$$\begin{cases} \psi_{t_2 t_1} = 0, & \text{если } t_1 \neq t_2 \begin{bmatrix} t_2 \in B_2,\ t_1 \in T_1 \\ t_2 \in T_2,\ t_1 \in B_1 \end{bmatrix} \\ \psi_{t_2 t_1} = 1, & \text{если } t_1 = t_2,\ t_2 \in B_2,\ t_1 \in B_1 \end{cases} \tag{7}$$

Для того, чтобы темы из B_{21} действительно унаследовали темы B_1, столбцы матрицы Φ_2, которые соответствуют темам B_2, предлагается проинициализировать столбцами матрицы Φ_1, которые соответствуют темам B_1.

Гипотеза 2. Чем больше вес межуровнего регуляризатора $\varkappa_1$, тем сильнее, жестче связь между первым и вторым уровнем:

- темы второго уровня неразнообразные, повторяют темы первого уровня;
- модель принимает структуру дерева.

Поэтому большие значения $\varkappa_1$ (> 1) нежелательны. Но и маленькие значения $\varkappa_1$ (< 0.1) могут привести к появлению большого количества избыточных связей, что тоже нежелательно.

Гипотеза 3. Порог для установки связи между темами первого и второго уровнями иерархической модели (который определяется по матрице Ψ) принимает значение, при котором у каждой темы второго уровня есть хотя бы одна родительская тема (ограничение сверху). При этом данный порог не допускает неинтерпретируемых связей (ограничение снизу). Действительно, если у темы второго уровня нет родительской темы, то полученная структура не удовлетворяет определению тематической иерархии. Однако, с другой стороны, это может означать, что темы первого уровня не покрывают всю тематику коллекции документов, и в этом случае необходимо пересмотреть первый уровень модели. Избыточные и неинтерпретируемые связи между уровнями модели только усложняют поиск документов для пользователей, что противоречит цели построения иерархической системы.

Гипотеза 4. Вес иерархического регуляризатора разреживания $\varkappa_2$ (4) выбирается таким образом, чтобы многодольный граф (соответствующий иерархической модели) был близок к дереву. Действительно, в таком графе пользователю проще ориентироваться и легче переходить в смежные области.

Поэтому предлагается устанавливать большой вес (≈ 10) для данного регуляризатора. При этом, следует следить за тем, чтобы модель не стала вырожденной: у темы со второго уровня нет родительской темы первого уровня. В связи с чем слишком большие веса устанавливать также не рекомендуется.

Таким образом, необходимо найти баланс между 2, 3 и 4 гипотезами, чтобы обеспечить разнообразные темы на втором уровне и получить модель со структурой, близкой к дереву.

2.3 Задача автоматического именования тем

Именование предметных тем $t \in S$ предлагается производить на основе модальности имен-кандидатов, и, благодаря разреженной вероятностной связи этой модальности с темами, подобрать для каждой темы наиболее подходящее имя. В данной работе в качестве модальности имён-кандидатов мы берём модальность тегов G. Предлагается формировать универсальный набор признаков $R_i(t, g)$, которые ранжируют теги $g \in G$ для каждой темы $t \in S$. Сравнение алгоритмов предлагается производить на основе асессорских оценок.

2.3.1 Признаки

Признаки предлагается извлекать из матриц Φ^g и Θ. При этом в качестве названий тем l-го уровня предлагается рассматривать только те теги $g \in G$, которые не были выбраны для именования тем $1, ..., l-1$ уровней иерархии. Для этого при формировании признаков для именовании тем l-го уровня в матрице Φ^g игнорируются строки, соответствующие тегам, уже выбранным в качестве названий для тем $1, ..., l-1$ уровней иерархии, а в матрице Θ рассматриваются только такие теги документов, которые отличны от названий тем предыдущих уровней иерархии $1, ..., l-1$.

Как отмечалось выше предлагается строить универсальный набор признаков, который не зависит от размерных характеристик задачи $n, |T|, |G|, |D|$.

При этом данные признаки предлагается измерять как по вероятностям $p(g|t)$ (то есть по Φ^g), так и по тегам документов G_d темы t (то есть по Θ). Значения признаков принимают значения от 0 до 1.

Первый признак $R_1(t, g)$ оценивает, насколько часто тег $g \in G$ встречается в теме $t \in S$: $p(g|t)$. Величины $p(g|t)$ зависят от $|G|$, поэтому функция ранжирования, построенная для одной задачи, будет неприменима к другой. Поэтому вместо условной вероятности $p(g|t)$ возьмем отношение:

$$R_1(t, g) = \left(\frac{p(g|t)}{p(g^*|t)}\right)^{\gamma_1} = \left(\frac{\phi_{gt}}{\phi_{g^*t}}\right)^{\gamma_1}, \tag{8}$$

где $g^* = \arg\max\limits_{g \in G} p(g|t)$.

Второй признак $R_2(t, g)$ оценивает долю документов темы $t \in S$, имеющих тег $g \in G$:

$$R_2(t, g) = \frac{r_2(g, t)}{r_2(g^*, t)}, \;\; r_2(g, t) = \frac{\sum\limits_{d \in D} \theta_{td} \left[g \in G_d\right]}{\sum\limits_{d \in D} \theta_{td}}, \tag{9}$$

где $g^* = \arg\max\limits_{g \in G} r_2(g, t)$. Тогда R_2 принимает вид:

$$R_2(t, g) = \left(\frac{\sum\limits_{d \in D} \theta_{td} \left[g \in G_d\right]}{\sum\limits_{d \in D} \theta_{td} \left[g^* \in G_d\right]}\right)^{\gamma_2},$$

Третий признак $R_3(t, g)$ оценивает вероятность выделения темы $t \in S$ для тега $g \in G$,:

$$R_3(t, g) = \left(\frac{p(t|g)}{p(t^*|g)}\right)^{\gamma_2}, \tag{10}$$

где $p(t|g) = \dfrac{p(g|t)p(t)}{p(g)}$ (формула Байеса), $t^* = \arg\max\limits_{t \in S} p(t|g)$. Воспользуемся формулами

$$p(t) = \sum_{d \in D} p(t|d)p(d), \;\; p(g) = \frac{n_g}{n}, \;\; p(d) = \frac{n_d}{n},$$

где n_g — число вхождений тега $g \in G$ в коллекцию D, n_d — длина документа $d \in D$ в тегах, n — длина коллекции D в тегах.

Тогда

$$p(t|g) = \frac{\phi_{gt}}{n_g} \sum_{d \in D} n_d \theta_{td}. \tag{11}$$

Третий признак принимает вид:

$$R_3(t, g) = \left(\frac{\phi_{gt} \sum\limits_{d \in D} n_d \theta_{td}}{\phi_{gt^*} \sum\limits_{d \in D} n_d \theta_{t^* d}} \right)^{\gamma_3}.$$

2.3.2 Экспертная оценка именования тем

Тема интерпретирума, если человек понимает о чем эта тема и может дать ей краткое именование.

Оценку качества именования тем различными алгоритмами предлагается провести на основе экспертной оценки по специальной методике.

Пусть темы $t_l \in T_l$ уровня $l \geq 1$ и всех вышележащих её уровней проименованы. Для именования предметных тем $t_{l+1} \in S_{l+1}$ $(l+1)$-го уровня каждому асессору предоставляется информация темы родительского уровня $t_l \in T_l$ и всех её дочерних тем $t_{l+1} \in S_{l+1}$ $(l+1)$-го уровня. Иными словами, эксперту по очереди показывают информацию по каждой теме с l-го уровня со всеми её дочерними темами с $(l+1)$-го уровня. Под информацией темы понимаются топ-слова всех её модальностей, за исключением модальности тегов. Также предоставляются названия тем l-го уровня. А для каждой темы $(l+1)$-го уровня предоставляется множество тегов $G_t \in G$. Задача эксперта заключается в том, чтобы каждой теме $(l+1)$-го уровня для каждого тега из списка поставить метку '++', '+' или '-': '++' — название абсолютно подходит для данной темы (то есть, лучше не придумаешь), '+' — название подходит для данной темы (то есть лучшее из имеющегося ссписка). При этом каждый знак для каждой темы можно выбрать несколько раз, если имена в равной степени подходят в качестве названия. В остальных случаях ставить знак '-', который стоит по умолчанию.

2.3.3 Оценка качества темы и ее названия

Пусть A – множество асессоров, $I(g,a) = [\text{асессор } a \text{ поставил } + \text{ для } g]$,

$g_t^* = \arg\max\limits_{g \in G_t} R_i(g,t)$ — имя, выбранное моделью для темы t.

Определим среднюю долю асессоров, согласных с именем, выбранным моделью:

$$MK = \frac{1}{|S|}\sum_{t\in S}\frac{1}{|A|}\sum_{a\in A} I(g_t^*, a). \tag{12}$$

Но МК зависит от того, насколько асессоры согласны между собой. Определим согласованность асессоров:

$$C = \frac{1}{|S|}\sum_{t\in S}\frac{1}{|A|}\sum_{a\in A}\frac{\sum\limits_{g\in G_t} I(g,a)\frac{1}{|A|-1}\sum\limits_{a'\in A\setminus a} I(g,a')}{\sum\limits_{g\in G_t} I(g,a)}. \tag{13}$$

3 Вычислительный эксперимент

3.1 Описание данных

Вычислительный эксперимент проводился на коллекции статей научно-популярного интернет-журнала ПостНаука. Коллекция состоит из 3404 документа и содержит модальности слов (19186 токенов), авторов (859 токенов), биграмм (11442), триграмм (464) и тегов (930). Теги к каждому документу были проставлены редакторами данного научного контента. Биграммы и триграммы были выделены алгоритмом, описанным в [3]. Известно, что n-граммы существенно повышают интерпретируемость тем.

Предобработка данных включает в себя нормализацию данных: перевод в нижний регистр, токенизацию и лемматизацию. Лемматизация была проведена морфологическим анализатором pymorphy2 [5]. Также были удалены редко и часто встречающиеся слова.

Модель строилась алгоритмом hARTM, реализованном в библиотеке BigARTM [4, 7].

3.2 Метрики качества моделей

Оценивание качества тематических моделей является нетривиальной проблемой. В отличие от задач классификации или регрессии здесь нет чёткого понятия «ошибки»

или «потери». Стандартные критерии качества кластеризации типа средних внутрикластерных или межкластерных расстояний или их отношений плохо подходят для оценивания «мягкой» совместной кластеризации документов и терминов.

Предполагается, что название каждой предметной темы соответствует одному тегу $t \in G$ для всех $t \in S$.

3.2.1 Перплексия

Наиболее распространённым критерием является перплексия. Это мера несоответствия или «удивлённости» модели $p(w|d)$ терминам w, наблюдаемым в документах d коллекции D, определяемая через логарифм правдоподобия:

$$\mathcal{P}(D) = \exp\left(\frac{1}{n}\sum_{d\in D}\sum_{w\in d} n_{dw} \ln p(w|d)\right),\ n = \sum_{d\in D}\sum_{w\in d} n_{dw}, \tag{14}$$

где n_{dw} — число терминов w в документе d.

Чем меньше эта величина, тем лучше модель p предсказывает появление терминов w в документах d коллекции D.

Интерпретация: если каждый документ генерируется из V равновероятных терминов (возможно, различных в разных документах), то перплексия сходится к V.

3.2.2 Разреженность

Разреженность модели измеряется долей $\mathcal{R}_{\Phi^m}$ и $\mathcal{R}_{\Theta}$ нулевых элементов в частях матриц Φ^m $m \in M$ и Θ, соответствующим предметным темам S.

3.2.3 Ошибка первого рода

False Positive Rate (FPR) — доля пар (d, t): тема t присутствует в d, но соответствующий ей тег $g = t$ не приписан документу d. Для формализации присутствия темы вводится порог k.

$$\mathrm{FPR}(k) = \frac{\sum\limits_{d}\sum\limits_{t\in S\setminus G_d}\left[\frac{\theta_{td}}{1-b(d)} \geqslant k\right]}{\sum\limits_{d}|S \setminus G_d|}. \tag{15}$$

3.2.4 Ошибка второго рода

False Negative Rate (FNR) — доля пар (d, t): тег g приписан документу d, а соответствующая ей тема $t = g$ в d отсутствут. Для формализации отсутствия темы вводится порог k.

$$\text{FNR}(k) = \frac{\sum\limits_{d} \sum\limits_{t \in G_d \cap S} [\frac{\theta_{td}}{1-b(d)} < k]}{\sum\limits_{d} |G_d \cap S|}. \tag{16}$$

3.3 Сравнение моделей

3.3.1 Построение первого уровня иерархии

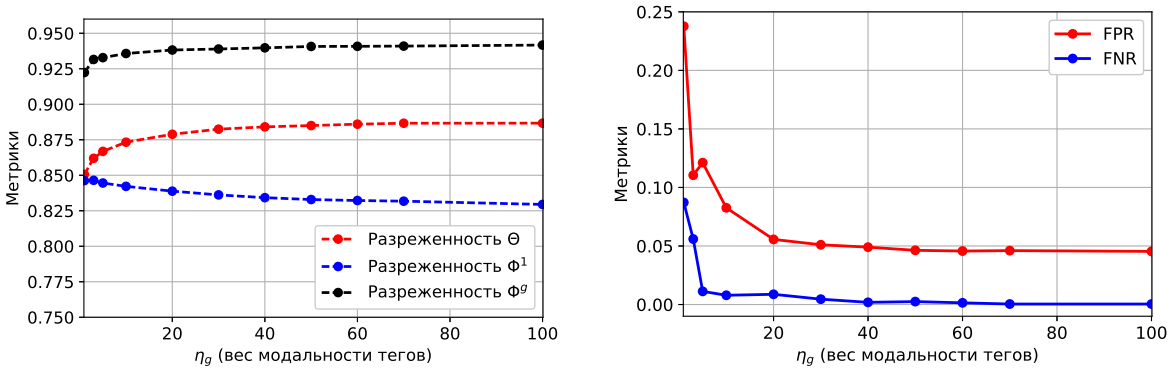
Редакторы контента ПостНаука выделили 20 тем для формирования первого уровня иерархической модели. При этом предложенные названия тем являются подмножеством множества тегов G: математика, технологии, физика, химия, земля, астрономия, биология, медицина, психология, экономика, история, политика, социология, культура, образование, язык, философия, религия, Россия, право. Таким образом, первый уровень иерархии строился с фиксированными числом тем — 21, среди которых 20 предметных тем и одна фоновая.

Для модальностей авторов и тегов были введены фиктивные автор и тег соответственно. Фоновая тема содержит только фиктивные токены и не содержит токены всех остальных авторов и тегов.

Модальности в модель включались поочередно в порядке их важности: слова и теги, авторы, биграммы, триграммы.

Теги

На начальном этапе в модель были включены две модальности: слова и теги. Для классификации документов по заданным темам выполнялась инициализация матрицы Φ как описано в разделе 2.1 по формуле (6). Качество классификации оценивалось с помощью ошибок первого $\text{FPR} = \text{FPR}(k^*)$ (15) и второго рода $\text{FNR} = \text{FNR}(k^*)$ (16), где $k^* = \arg\min\limits_k \text{FPR}(k) + \text{FNR}(k)$. Как и предполагалось, качество классификации зависит от веса модальности тегов η_g в уравнении (1). На рис. 1(a)-(b) видно, что с определенного значения η_g кривые, отражающие зависимость значений метрик качеств от η_g, выходят на насыщение. В качестве η_g было



(a) (b)

Рис. 1: Графики зависимости значений метрик качеств от веса модальности тегов G для первого уровня иерархической модели, в которую включены модальности слов и тегов.

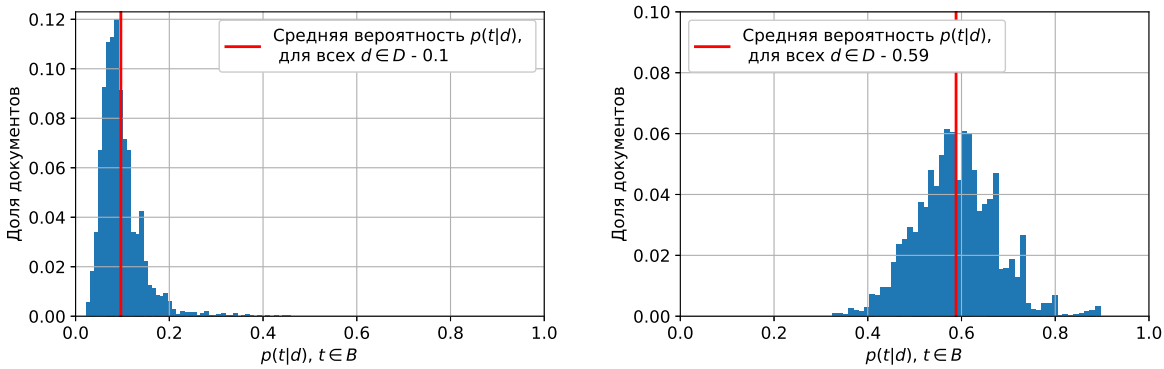
выбрано значение 60, при котором FPR $= 0.02$, FNR $= 0.00$, что говорит о хорошем качестве классификации.

Фоновая тема

Для того, чтобы фоновая тема $t \in B_1$ действительно содержала все слова общей лексики, а предметные темы $t \in S_1$ только специфичные для них токены, предлагается в модель включать регуляризатор сглаживания документов по фоновым темам B_1. В процессе выбора веса данного регуляризатора строились гистограммы, отражающие распределение документов по доле присутствия фоновой темы $t \in B_1$ в них (рис. 2). В результате экспериментов был выбран вес $\tau = 500$, при котором модель в среднем в каждом документе $d \in D$ 60% токенов относит к фоновой теме $t \in B_1$. На рис. 3 представлены примеры топ слов фоновых тем при разном весе регуляризатора сглаживания документов по фоновой теме.

Выбор веса регуляризаторов. Выбор веса регуляризаторов происходил аналогично технике, описанной выше для выбора веса для модальности тегов.

Регуляризатор включался в модель с таким весом, при котором распределние модальностей в темах интерпретируемо и значения метрик модели, по крайней мере, ухудшаются незначительно. Так как решается задача классификации, то определящую роль при выборе веса регуляризатора играют ошибки первого и второго рода (FPR и FNR). В таблице 1 видно, что в результате добавления в модель регуля-



(a) $\tau = 30$ (b) $\tau = 500$

Рис. 2: Распределение документов по доле присутствия фоновой темы $t \in B_1$ в них при разном весе регуляризатора сглаживания документов по фоновой теме τ для первого уровня иерархии.

книга 0.016,знать 0.009,сторона 0.007,работа 0.007,важный 0.006,большой 0.006,проблема 0.006,вопрос 0.006,исследование 0.006,работать 0.006,говорить 0.006,разный 0.006,существовать 0.006,должный 0.005,хороший 0.005,автор 0.005,образ 0.005,общий 0.005,наука 0.004,хотеть 0.004,задача 0.004,метр 0.004,думать 0.004,ситуация 0.004,дать 0.004,развитие 0.004,последний 0.004,стать 0.004,случай 0.004,результат 0.004,тема 0.004,изучать 0.004,понять 0.004,процесс 0.004,понимание 0.004,взгляд 0.004,идея 0.004,понимать 0.003,написать 0.003,учёный 0.003,жизнь 0.003,некоторый 0.003,давать 0.003,многие 0.003,видеть 0.003,область 0.003,далёкий 0.003,являться 0.003,появиться 0.003,смысл	говорить 0.005,являться 0.005,большой 0.005,образ 0.005,должный 0.005,существовать 0.005,стать 0.004,работа 0.004,сторона 0.004,книга 0.004,знать 0.004,называть 0.004,вопрос 0.004,важный 0.004,жизнь 0.004,происходить 0.004,некоторый 0.004,проблема 0.004,система 0.003,исследование 0.003,разный 0.003,связать 0.003,работать 0.003,дать 0.003,результат 0.003,процесс 0.003,хороший 0.003,видеть 0.003,случай 0.003,вид 0.003,друг 0.003,поскольку 0.003,отношение 0.003,наука 0.002,новый 0.002,находиться 0.002,число 0.002,оказаться 0.002,идея 0.002,имя 0.002,использовать 0.002,последний 0.002,история 0.002,метр 0.002,область 0.002,начинать 0.002,часть 0.002,современный 0.002,принцип 0.002,ситуация 0.002

(a) $\tau = 30$ (b) $\tau = 500$

Рис. 3: Топ слова фоновой темы $t \in B$ с их вероятностями принадлежности к t при разном весе регуляризатора сглаживания документов по фоновой теме τ для первого уровня иерархии.

ризатора разреживания матрицы Θ значения всех метрик улучшились. Модальности авторов, биграмм и триграмм удалось добавить в модель, практически не ухудшив значения метрик. Модальности ценны для построения последующих уровней иерархии.

Регуляризатор разреживания матрицы Θ

Модель лучше интерпретируема, если каждый документ содержит небольшое число тем — это означает, что матрица Θ должна быть разреженной. Для этого следует применять регуляризатор разреживания матрицы Θ для предметных тем.

Логично предположить, что в коллекции практически нет междисциплинарных

Номер модели	**Регуляризотор**	**Вес**	$\mathcal{R}_\Phi$	$\mathcal{R}_\Theta$	$\mathcal{P}$	FPR	FNR
1	Модальность тегов	60	0.832	0.886	4074	0.046	0.002
2	Сглаживание $t \in B$ в Θ	500	**0.841**	**0.893**	**4347**	**0.046**	**0.000**
3	Модальность авторов	100	0.837	**0.896**	**4243**	**0.043**	0.001
4	Модальность биграмм	1	0.836	**0.898**	**3001**	**0.042**	0.001
5	Модальность триграмм	1	**0.836**	**0.898**	**2925**	0.043	**0.000**
6	Разреживание $t \in S$ в Θ	-30	0.830	**0.927**	2970	**0.020**	**0.000**
7	Декоррелирование слов	0.2	**0.921**	0.925	3632	**0.020**	**0.000**
8	Декоррелирование тригр	0.1	**0.921**	**0.925**	**3632**	**0.020**	**0.000**

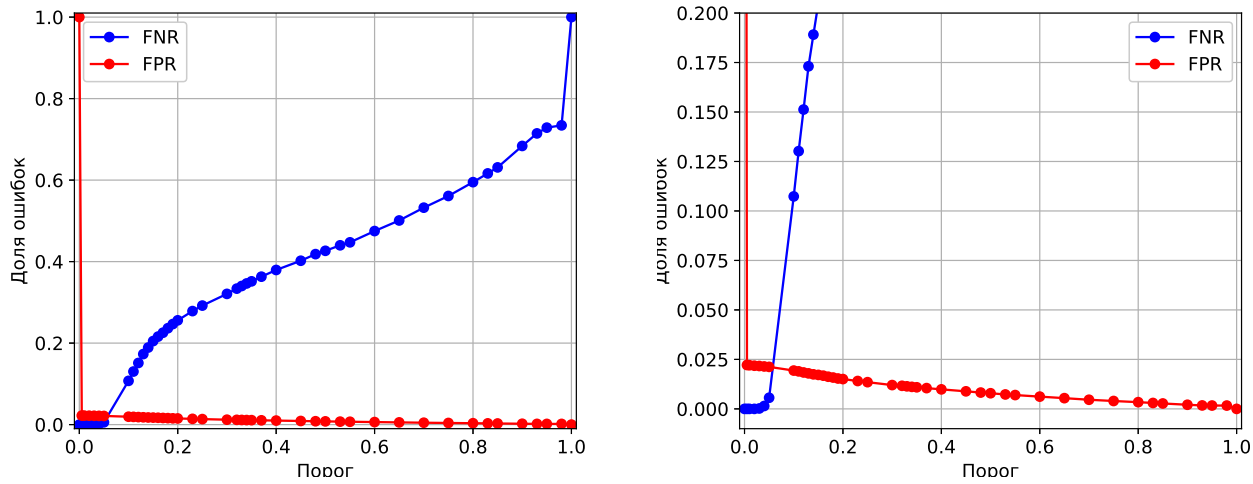
Таблица 1: Значения метрик для различных моделей (модели пронумерованы в порядке добавления регуляризоторов).

документов, затрагивающих несколько крупных областей знания, то есть каждый документ содержит одну тему первого уровня, — разреженность матрицы Θ составляет $1 - \frac{1}{|S|} = 0.95$. Значение 0.9 уже означает, что в среднем каждый документ затрагивает два больших раздела науки, что вряд ли реалистично.

При разреживании матрицы Θ важно следить еще за одним параметром для каждой темы: количество документов данной темы $t \in S_1$, имеющих такой тег $g \in G$, который совпадает с названием данной темы $g = t$. Это число должно быть не меньше числа документов, имеющих данный тег g. При разреженности 0.93 данное условие нарушается.

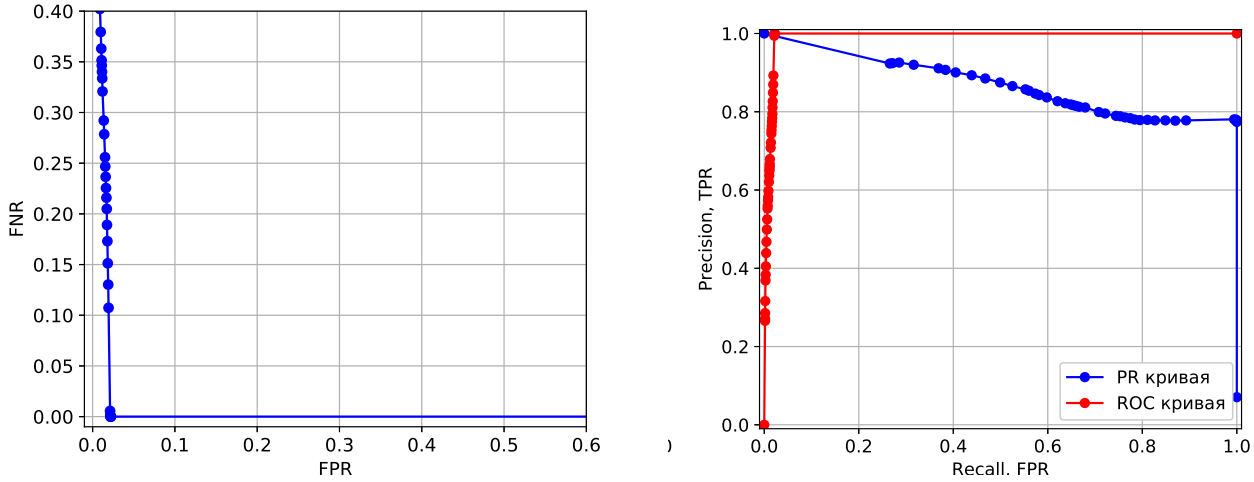
Ошибки первого и второго рода

Для каждой модели исследовались ошибки первого FPR (15) и второго рода FNR (16). На рис. 4 и 5 представлены результаты для итоговой 8-ой модели (табл. 1). На рис. 5(a) видно, что есть интервал значений порога, при котором модель 8 практически не допускает ошибок второго рода FNR, приэтом и значения ошибок первого рода FPR невелико (≈ 0.025). Согласно рис. 4 такие значения ошибок достигаются при значениях порога k от 0 до 0.05. По графикам Precision-Recall и ROC-кривых (рис. 5(b)) можно сделать вывод, что достигнуто довольно неплохое качество классификации.



(a) (b)

Рис. 4: Графики зависимости ошибок первого FPR(k) и второго FNR(k) рода для модели 8 (табл. 1).



(a) (b)

Рис. 5: График зависимости FNR(k) от FPR(k) (a), графики Precision-Recall и ROC кривых (b) для модели 8 (табл. 1).

3.3.2 Построение второго уровня иерархии

Второй уровень иерархии строился с $|S_2| = 60$ предметными и $|B_2| = 2$ фоновыми темами. При этом одна из фоновых тем $t_{21} \in B_2$ наследует фоновую тему первого уровня $t_1 \in B_1$, а вторая тема $t_{22} \in B_2$ является фоновой темой плоской модели,

описывающей второй уровень иерархии. Для достижения такого результата был введен регуляризатор разреживания матрицы Ψ, которая является подматрицей Θ_2, таким образом, чтобы выполнялись соотношения (7). Столбец Φ_2, соответствующий теме t_{21}, был проинициализирован столбцом Φ_1, соответствующим теме t_1. Данная техника позволила добиться того, чтобы $t_1 \in B$ не являлась родительской темой ни для одной предметной темы второго уровня.

Выбор веса регуляризаторов происходил по той же технике, что и для первого уровня иерархии. Но так как данная задача не является задачей классификацией, то оценивать ошибки $\mathrm{FPR}(k)$ и $\mathrm{FNR}(k)$ нет возможности. Поэтому модели оценивались по $\mathcal{R}_{\Psi_2^1}$, $\mathcal{R}_{\Theta_2}$ и $\mathcal{P}$.

Важную роль при построении второго уровня иерархии играют межуровневый регуляризатор и иерархический регуляризатор разреживания, так как они влияют на связь тем второго уровня с темами первого уровня. Также они определяют структуру иерархии, то есть определяют насколько многодольный граф близок к дереву. На рис. 6 представлена зависимость среднего количества родителей тем второго уровня от веса иерархического регуляризатора разреживания $\varkappa_2$ при различных значениях веса межуровнего регуляризатора $\varkappa_1$. Видно, что когда $\varkappa_2$ принимает значение $\approx 10^5$ модель вырождается в дерево при любом значении $\varkappa_1$. При $\varkappa_1 \approx 10$ модель также принимает структуру дерева при любом значении $\varkappa_1$. Такое поведение подтверждает гипотезы 2 и 4.

3.4 Автоматическое именование тем

Качество именования тем оценивалось на основе асессорских оценок. Всего было 3 эксперта. Для каждой темы t второго уровня предлагалось $G_t = 20$ возможных названий из множества тегов G. Согласованность экспертов C составила 0.54.

В таблице 2 представлены значения средних долей асессоров, согласных с именем, выбранным различными моделями. При согласованности асессоров $C = 0.54$ результаты ранжировок признаков R_1 и R_2 дали довольно высокие значения MK.

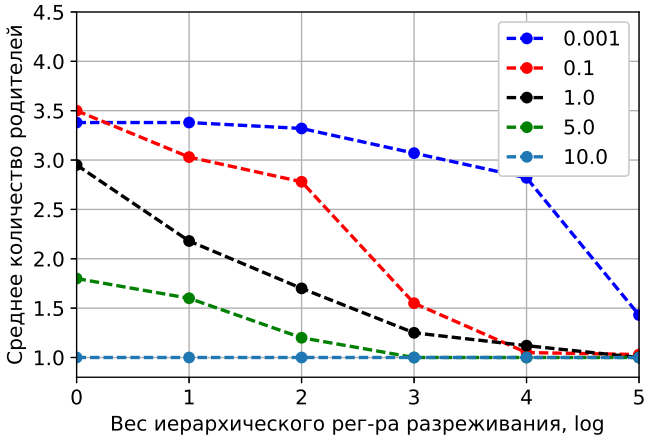


Рис. 6: График зависимости среднего количества родителей тем второго уровня от веса иерархического регуляризатора разреживания при различных значениях веса межуровнего регуляризатора.

Признак	**MK**
R_1	0.6
R_2	0.62
R_3	0.32
R_1R_2	0.62

Таблица 2: Средняя доля асессоров, согласных с именем, выбранным моделью, для второго уровня иерархии при согласованности асессоров $C = 0.54$.

3.5 Выводы и рекомендации

1. Модальности необходимо включать в модель последовательно, чтобы была возможность оценить степень влияния каждой модальности по отдельности.

2. Важно определить последовательность включения модальностей в модель, что зависит от специфики задачи. Вначале необходимо включать модальности, которые определяют модель, а последующие модальности корректируют уже неплохо построенную модель. Для нашей задачи важными являются модальности слов, тегов и авторов, а для уточнения и корректировки используются модальности биграмм и триграмм.

3. Необходимо понять, каким образом можно учесть ограничения, накладываемые на модель (если они есть). В частности, на модель можно сильно повлиять, сделав специфическую инициализацию матрицы Φ.

4. Регуляризаторы декоррелирования тем по различным модальностям следует включать в модель после добавления всех модальностей. Если после добавления модальности m_1 в модель сразу же применить регуляризатор декоррелирования тем по модальности m_1, то добавление модальности m_2 может привести к вырожденной модели.

5. К фоновой теме следует применить регуляризатор сглаживания. Это приведет к тому, что предметные темы будут состоять только из специфичных токенов, так как токены общей лексики будут содержаться в фоновой теме. Данный регуляризатор следует вводить после добавления важных модальностей, также его вес можно корректировать и после дообавления последующих модальностей.

6. Модель лучше интерпретируема, если каждый документ содержит небольшое число тем – это означает, что матрица Θ должна быть разреженной. Для этого следует применять регуляризатор разреживания матрицы Θ для предметных тем.

7. При построении второго уровня иерархии важно найти баланс между значениями весов межуровневого иерархического регуляризатора и иерархического регуляризатора разреживания для того, чтобы темы были разнообразными, а структура иерархии была близка к дереву. Выполнение данных условий упростят пользователям осуществлять поиск необходимых документов в коллекции.

8. Наличие модальности имён-кандидатов позволяет осуществить автоматическое именование тем с хорошим качеством.

Заключение

В работе предложен метод послойного построения иерархии коллекций научно-популярных текстов. При решении предполагалось, что документы тегированы, то есть каждому документу редакторами ресурса приписано некоторое количество ключевых слов или фраз. При этом для первого уровня иерархии известно количество тем и их названия, которые являются подмножеством множества тегов. Таким образом, задача построения тематической модели первого уровня иерархии свелась к классификации документов по заданным темам. Предложенный метод для решения данной задачи позволил классифицировать документы с хорошей точностью, ошибки первого и второго рода составили 0.02 и 0.00 соответственно на коллекции ПостНаука. Для остальных уровней иерархии предложен алгоритм автоматического именования тем. Эксперименты показали, что алгоритм выдает названия тем, которые хорошо согласованы с именами, выбранными асессорами.

Список литературы

[1] David M. Blei, Thomas L. Griffiths, Michael I. Jordan, and Joshua B. Tenenbaum. Hierarchical topic models and the nested chinese restaurant process. In *Advances in Neural Information Processing Systems*, page 2003. MIT Press, 2004.

[2] Vorontsov K. V. Chirkova N. A. Additive regularization for hierarchical multimodal topic modeling. *JMachine Learning and Data Analysis*, 2, 2016.

[3] Ahmed El-Kishky, Yanglei Song, Chi Wang, Clare R. Voss, and Jiawei Han. Scalable topical phrase mining from text corpora. *Proc. VLDB Endow.*, 8(3):305–316, November 2014.

[4] Oleksandr Frei and Murat Apishev. *Parallel Non-blocking Deterministic Algorithm for Online Topic Modeling*, pages 132–144. Springer International Publishing, Cham, 2017.

[5] Mikhail Korobov. Morphological analyzer and generator for russian and ukrainian languages. In Mikhail Yu. Khachay, Natalia Konstantinova, Alexander Panchenko,

Dmitry I. Ignatov, and Valeri G. Labunets, editors, *Analysis of Images, Social Networks and Texts*, volume 542 of *Communications in Computer and Information Science*, pages 320–332. Springer International Publishing, 2015.

[6] David Mimno, Wei Li, and Andrew McCallum. Mixtures of hierarchical topics with pachinko allocation. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, pages 633–640, New York, NY, USA, 2007. ACM.

[7] Konstantin Vorontsov, Oleksandr Frei, Murat Apishev, Peter Romov, and Marina Dudarenko. Bigartm: Open source library for regularized multimodal topic modeling of large collections. In *Analysis of Images, Social Networks and Texts - 4th International Conference, AIST 2015, Yekaterinburg, Russia, April 9-11, 2015, Revised Selected Papers*, pages 370–381, 2015.

[8] Konstantin Vorontsov, Oleksandr Frei, Murat Apishev, Peter Romov, Marina Suvorova, and Anastasia Yanina. Non-bayesian additive regularization for multimodal topic modeling of large collections. pages 29–37, 2015.

[9] Konstantin Vorontsov and Anna Potapenko. Tutorial on probabilistic topic modeling: Additive regularization for stochastic matrix factorization. pages 29–46, 2014.

[10] Konstantin Vorontsov, Anna Potapenko, and Alexander Plavin. *Additive Regularization of Topic Models for Topic Selection and Sparse Factorization*. Springer International Publishing, Cham, 2015.

[11] C. Wang, M. Danilevsky, J. Liu, N. Desai, H. Ji, and J. Han. Constructing topical hierarchies in heterogeneous information networks. In *2013 IEEE 13th International Conference on Data Mining*, pages 767–776, Dec 2013.

[12] Chi Wang, Marina Danilevsky, Nihit Desai, Yinan Zhang, Phuong Nguyen, Thrivikrama Taula, and Jiawei Han. A phrase mining framework for recursive construction of a topical hierarchy. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, pages 437–445, New York, NY, USA, 2013. ACM.

[13] Chi Wang, Xueqing Liu, Yanglei Song, and Jiawei Han. Scalable and robust construction of topical hierarchies. *CoRR*, abs/1403.3460, 2014.

[14] Chi Wang, Xueqing Liu, Yanglei Song, and Jiawei Han. Towards interactive construction of topical hierarchy: A recursive tensor decomposition approach. In *Proc. 2015 ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'15)*. ACM – Association for Computing Machinery, August 2015.

[15] Chi Wang, Xueqing Liu, Yanglei Song, and Jiawei Han. Towards interactive construction of topical hierarchy: A recursive tensor decomposition approach. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 1225–1234, New York, NY, USA, 2015. ACM.

[16] Elias Zavitsanos, Georgios Paliouras, and George A. Vouros. Non-parametric estimation of topic hierarchies from texts with hierarchical dirichlet processes. *Journal of Machine Learning Research*, 12:2749–2775, 2011.