

Министерство образования и науки Российской Федерации

Федеральное государственное автономное образовательное  
учреждение высшего профессионального образования

«Московский физико-технический институт  
(государственный университет)»

Факультет управления и прикладной математики

Кафедра «Интеллектуальные Системы»

## **Построение и оценка качества гетерогенных иерархических тематических моделей**

Выпускная квалификационная работа  
(бакалаврская работа)

Направление подготовки: 03.03.01 Прикладные математика и физика

Выполнил:

студент 474 группы \_\_\_\_\_ Селезнева Мария  
Сергеевна

Научный руководитель:

к.ф.-м.н. \_\_\_\_\_ Воронцов Константин  
Вячеславович

Москва 2018

# Оглавление

<b>1</b>	<b>Введение</b>	<b>4</b>
<b>2</b>	<b>Обзор литературы</b>	<b>6</b>
2.1	Постановка задачи тематического моделирования . . .	6
2.2	Плоские тематические модели . . . . .	7
2.2.1	Вероятностный латентный семантический анализ	7
2.2.2	Байесовский подход . . . . .	8
2.2.3	Аддитивная регуляризация тематических моделей	8
2.2.4	Мультимодальные тематические модели . . . .	9
2.3	Иерархические тематические модели . . . . .	10
2.3.1	Обобщения LDA для тематических иерархий . .	10
2.3.2	Иерархическая модель ARTM . . . . .	10
2.4	Метрики качества тематических моделей . . . . .	11
<b>3</b>	<b>Измерение качества тематических иерархий</b>	<b>14</b>
3.1	Метрики качества ребер иерархии . . . . .	14
3.1.1	Метрики на основе лингвистической близости .	14
3.1.2	Метрики на основе вероятностной близости . .	15
3.2	Ассессорская разметка ребер иерархии . . . . .	15
3.2.1	Описанные данных и моделей . . . . .	16
3.2.2	Постановка задания для ассессоров . . . . .	16
3.2.3	Контроль качества . . . . .	17
3.2.4	Результаты . . . . .	17
3.3	Сравнение метрик с ассессорскими оценками . . . . .	18
3.4	Метрики качества тематических иерархий . . . . .	19
<b>4</b>	<b>Агрегирование гетерогенных текстовых коллекций в тематической иерархии</b>	<b>20</b>
4.1	Постановка задачи . . . . .	20
4.2	Описание данных . . . . .	20
4.3	Базовый алгоритм . . . . .	21
4.4	Предлагаемый алгоритм . . . . .	22

4.4.1	Фильтрация новой коллекции . . . . .	22
4.4.2	Дополнение модели отфильтрованной коллекцией	23
4.5	Сравнение алгоритмов . . . . .	23
<b>5</b>	<b>Заключение</b>	<b>24</b>

# Todo list

# Глава 1

## Введение

Вероятностное тематическое моделирование — это раздел машинного обучения, решающий задачу поиска тем в коллекции документов. Тематическая модель определяет, к каким темам относится каждый документ и какие слова образуют темы [1]. Базовыми подходами построения тематических моделей являются PLSA [2] и LDA [3]. В работах [4, 5, 6] предложены модификации данных подходов, учитывающие специфику конкретных задач. Аддитивная регуляризация тематических моделей (ARTM) [7, 8, 9, 10, 11] позволяет комбинировать упомянутые модели, интерпретируя их как регуляризаторы в PLSA.

В больших текстовых коллекциях темы часто образуют иерархии, в которых каждая тема делится на более специфичные подтемы. Моделью такой ситуации являются тематические иерархии. Они удобны для навигации пользователей по коллекциям, поэтому являются подходящей моделью для агрегирования контента.

Общепринятого определения и подхода к построению иерархических тематических моделей не существует. В модели иерархического LDA (hLDA) [12] темы образуют дерево. С другой стороны, модель иерархического распределения патинко (hPAM) [13] и модель иерархического ARTM (hARTM) [14] представляют собой направленный ациклический многодольный граф, что лучше соответствует реальным отношениям между темами в мультидисциплинарных научных и научно-популярных статьях.

Модель hARTM — это развитие идеи аддитивной регуляризации тематических моделей для задачи построения тематических иерархий. Она позволяет применять регуляризацию как к уровням иерархии для комбинирования любых тематических моделей, так и к самой иерархии для контроля разреженности отношения «родитель-ребенок».

Цель данной работы — построение иерархической тематической модели по нескольким источникам в рамках модели hARTM. Постро-

ение модели по объединенной коллекции источников, различных по объему и тематической структуре, не решает поставленной задачи, так как темы, уникальные для меньшего из источников, теряются. В работе предлагается дополнять существующую модель одного источника выборками документов из новых источников. Дополнение происходит в два этапа: сначала выбираются документы нового источника, наиболее подходящие для добавления в коллекцию, затем выбранные документы добавляются в коллекцию и строится дополненная модель. При этом в силу стратегии инициализации и регуляризации модели существующие темы сохраняются, а добавленные документы уточняют соответствующие им темы на первом уровне иерархии. На втором уровне добавленные документы могут порождать подтемы, характерные только для нового источника.

В работах [15, 16] рассмотрено построение иерархических тематических моделей гетерогенных источников. Однако, задача последовательного достроения модели затронута в работах, не использующих подход вероятностного тематического моделирования [17]. Предложенный в данной работе подход выигрывает у построения по объединенной коллекции по качеству решения и скорости.

Для проведения экспериментов используется BigARTM — библиотека для тематического моделирования с открытым исходным кодом [18, 19].

## Глава 2

# Обзор литературы

### 2.1 Постановка задачи тематического моделирования

В вероятностном тематическом моделировании коллекция документов рассматривается как множество троек  $(d, w, t)$ , выбранных случайно и независимо из дискретного распределения  $p(d, w, t)$ , заданного на конечном множестве  $D \times W \times T$ . Здесь  $D$  – множество документов коллекции,  $W$  – словарь,  $T$  – множество тем. Документы  $d \in D$  и токены  $w \in W$  являются наблюдаемыми переменными, а тема  $t \in T$  является латентной (скрытой) переменной.

Построить тематическую модель коллекции документов  $D$  — значит найти распределения  $p(w|t)$  для всех тем  $t \in T$  и распределения  $p(t|d)$  для всех документов  $d \in D$ . Известными при это являются распределения  $p(w|d)$  терминов (токенов) в документах коллекции.

Предполагается, что порядок токенов  $w \in W$  в документе  $d \in D$  не важен (гипотеза “мешка слов”), что позволяет представить коллекцию в виде матрицы  $[n_{dw}]_{D \times W}$ , где  $n_{dw}$  — число вхождений  $w$  в  $d$ . Также коллекцию можно представить в виде матрицы эмпирических оценок вероятности встретить токен в документе  $[p_{wd}]_{W \times D}$ :

$$p_{wd} = \hat{p}(w|d) = \frac{n_{dw}}{n_d},$$

где  $n_d = \sum_{w \in W} n_{dw}$  — число слов в документе  $d$ .

Кроме того, принимается гипотеза условной независимости:

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

То есть вероятность появления токена в некоторой теме не зависит от того, в каком документе встретился этот токен.

## 2.2 Плоские тематические модели

### 2.2.1 Вероятностный латентный семантический анализ

При сделанных предположениях плоская (одноуровневая) тематическая модель описывается формулой

$$p(w|d) \approx \sum_{t \in T} p(w|t)p(t|d) \quad d \in D, w \in W,$$

которая следует из определения условной вероятности и формулы полной вероятности. Пусть число тем  $|T|$  много меньше числа документов  $|D|$  и числа терминов  $|W|$ . Тогда представим задачу в виде факторизации матрицы  $F = [p_{wd}]_{W \times D}$ :

$$F \approx \Phi \Theta.$$

Параметры модели — матрицы  $\Phi = [\phi_{wt}]_{W \times T}$ ,  $\phi_{wt} = p(w|t)$  (вероятности токенов в темах) и  $\Theta = [\theta_{td}]_{T \times D}$ ,  $\theta_{td} = p(t|d)$  (вероятности тем в документах).

Первая модель, использующая описанный подход, называется вероятностный латентный семантический анализ (PLSA) [2]. В ней матрицы  $\Phi$  и  $\Theta$  находятся с помощью максимизации логарифма правдоподобия:

$$\sum_{d \in D} \sum_{w \in W} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td} \rightarrow \max_{\Phi, \Theta}$$

$$\text{при условиях } \phi_{wt} \geq 0, \theta_{td} \geq 0, \sum_{w \in W} \phi_{wt} = 1, \sum_{t \in T} \theta_{td} = 1.$$

Эта оптимизационная задача решается ЕМ-алгоритмом:

Е-шаг:

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

М-шаг:

$$\begin{aligned} \phi_{wt} &= \frac{n_{wt}}{\sum_{v \in W} n_{vt}}; & n_{wt} &= \sum_{d \in D} n_{dw} p_{tdw}; \\ \theta_{td} &= \frac{n_{td}}{\sum_{s \in T} n_{sd}}; & n_{td} &= \sum_{w \in W} n_{dw} p_{tdw}. \end{aligned}$$



Задача является некорректно поставленной, так как она допускает бесконечное множество решений. Действительно, для любой матрицы  $S$  ранга  $|T|$  имеем

$$\Phi\Theta = (\Phi S)(S^{-1}\Theta).$$

## 2.2.2 Байесовский подход

Байесовский подход к задаче тематического моделирования рассматривает оптимизационную задачу как максимизацию апостериорной вероятности. Тогда параметры модели генерируются из некоторых априорных распределений.

Наиболее применимой моделью является латентное размещение Дирихле (LDA) [3], в которой вектора  $\phi_t$  и  $\theta_d$  генерируются из распределений Дирихле. Кроме того, было предложено множество обобщений LDA, учитывающих дополнительные требования к модели [4, 5, 6].

## 2.2.3 Аддитивная регуляризация тематических моделей

Априорным распределениям параметров тематической модели соответствуют регуляризаторы в задаче оптимизации. Аддитивная регуляризация тематических моделей (модель ARTM) [10] позволяет одновременно учитывать множество дополнительных требований к модели.

В ARTM матрицы  $\Phi$  и  $\Theta$  находятся с помощью максимизации суммы логарифма правдоподобия и регуляризаторов  $R_i$  с неотрицательными коэффициентами регуляризации  $\tau_i$ :

$$\sum_{d \in D} \sum_{w \in W} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td} + \sum_i \tau_i R_i(\Phi, \Theta) \rightarrow \max_{\Phi, \Theta}$$

$$\text{при условиях } \phi_{wt} \geq 0, \theta_{td} \geq 0, \sum_{w \in W} \phi_{wt} = 1, \sum_{t \in T} \theta_{td} = 1.$$

Как показано в [10], EM-алгоритм для этой задачи имеет следующий вид:

Е-шаг:

$$p(t|d, w) = p_{tdw} = \text{norm}_{t \in T}(\phi_{wt} \theta_{td}).$$

М-шаг:

$$\phi_{wt} = \text{norm}_{w \in W} \left( n_{wt} + \phi_{wt} \sum_i \tau_i \frac{\partial R_i}{\partial \phi_{wt}} \right); \quad n_{wt} = \sum_{d \in D} n_{dw} p_{tdw};$$

$$\theta_{td} = \text{norm}_{t \in T} \left( n_{td} + \theta_{td} \sum_i \tau_i \frac{\partial R_i}{\partial \theta_{td}} \right); \quad n_{td} = \sum_{w \in d} n_{dw} p_{tdw},$$

$$\text{где введен оператор } \text{norm}_{t \in T}(x_t) = \frac{\max\{x_t, 0\}}{\sum_{s \in T} \max\{x_s, 0\}}.$$

Тогда модель PLSA эквивалентна модели ARTM в случае  $R(\Phi, \Theta) = \sum_i \tau_i R_i(\Phi, \Theta) = 0$ .

Регуляризаторы ARTM не обязаны иметь вероятностные интерпретации и могут учитывать любые особенности модели. Приведем описание регуляризаторов, используемых в данной работе.

#### **Сглаживание и разреживание:**

Регуляризатор сглаживания вводит в модель требование, чтобы столбцы  $\phi_t$  и  $\theta_d$  были близки к заданным распределениям  $\beta_t = [\beta_{wt}]_{w \in W}$  и  $\alpha_d = [\alpha_{td}]_{t \in T}$  в смысле дивергенции Кульбака-Лейблера [10]:

$$R(\Phi, \Theta) = \beta_0 \sum_{m \in M} \sum_{t \in T} \sum_{w \in W^m} \beta_{wt} \ln \phi_{wt} + \alpha_0 \sum_{d \in D} \sum_{t \in T} \alpha_{td} \ln \theta_{td} \rightarrow \max_{\Phi, \Theta},$$

где  $\beta_0$  и  $\alpha_0$  — заданные положительные коэффициенты. Введение этого регуляризатора, как показано в [10], эквивалентно модели LDA.

Регуляризатор разреживания имеет такой же вид, но коэффициенты  $\beta_0$  и  $\alpha_0$  отрицательны. Он способствует обращению значительной части вероятностей  $\phi_{wt}$  и  $\theta_{td}$  в ноль, что соответствует естественному предположению о том, что каждый документ  $d$  и каждый токен  $w$  связаны лишь с небольшим числом тем  $t$ .

#### **Декоррелирование:**

Регуляризатор декоррелирования минимизирует ковариации между столбцами  $\phi_t$ , что способствует увеличению различности тем модели [10]:

$$R(\Phi) = -\frac{\tau}{2} \sum_{t \in T} \sum_{s \in T \setminus t} \text{cov}(\phi_t, \phi_s) \rightarrow \max_{\Phi, \Theta}, \quad \text{cov}(\phi_t, \phi_s) = \sum_{w \in W} \phi_{wt} \phi_{ws}.$$

## **2.2.4 Мультимодальные тематические модели**

Документы  $d \in D$  могут содержать не только текст, но и другие элементы, такие как тэги, ссылки, имена авторов и т.д. Такие типы элементов называются модальностями. Пусть  $M$  — множество модальностей. Каждой  $m \in M$  соответствует отдельный словарь (множество токенов)  $W^m$ , причем  $W = \bigsqcup_{m \in M} W^m$ .

Пусть  $F^m = [p_{wd}]_{W^m \times D}$ ,  $m \in M$  — матрицы наблюдаемых вероятностей для каждой модальности, а  $\Phi^m$  — соответствующие матрицы скрытых вероятностей  $p(w|t)$ . Определим  $F$  и  $\Phi$  как объединения

строк  $F^m$  и  $\Phi^m$ ,  $m \in M$  соответственно. Тогда получим задачу факторизации  $F \approx \Phi\Theta$  для мультимодальной тематической модели.

При построении мультимодальной тематической модели максимизируется взвешенная сумма логарифмов правдоподобия для всех модальностей  $m \in M$  и регуляризаторов  $R_i$  [10]:

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

$$\text{при условиях } \phi_{wt} \geq 0, \theta_{td} \geq 0, \sum_{w \in W} \phi_{wt} = 1, \sum_{t \in T} \theta_{td} = 1.$$

## 2.3 Иерархические тематические модели

Тематическая иерархия представляет собой многоуровневый граф, где каждый уровень — это плоская тематическая модель. Для ее построения необходимо не только строить тематические модели уровней, но и устанавливать связи родитель-ребенок между темами соседних уровней. При этом общепринятого определения и подхода к построению иерархических тематических моделей не существует.

### 2.3.1 Обобщения LDA для тематических иерархий

Многие иерархические тематические модели были разработаны как обобщения модели LDA. Первой такой моделью являлось иерархическое LDA (hLDA) [12]. В нем темы образуют дерево, то есть каждая подтема (тема-ребенок) имеет только одну тему-родителя.

С другой стороны, модель иерархического распределения патинко (hPAM) [13] представляет собой направленный ациклический многодольный граф. Модель иерархического процесса Дирихле (hHDP) [20] также является многодольным графом и дополнительно обеспечивает возможность оценивать количество уровней и количество тем на каждом уровне иерархии.

Существуют масштабируемые модели тематических иерархий, которые подходят для больших наборов данных, например [21, 22].

### 2.3.2 Иерархическая модель ARTM

В иерархической модели ARTM (hARTM) [14] для связи уровней иерархии вводятся специальные межуровневые регуляризаторы. При этом иерархия является многодольным ациклическим графом.

Пусть построено  $l \geq 1$  уровней тематической иерархии, параметры  $l$ -того уровня — матрицы  $\Phi^l$ ,  $\Theta^l$ ,  $A$  — множество тем  $l$ -го уровня. Построим  $(l + 1)$ -ый уровень с параметрами  $\Phi$ ,  $\Theta$  и множеством тем  $T$ .

Будем моделировать распределение токенов по темам  $l$ -того уровня как смесь распределений по темам  $(l + 1)$ -го уровня [14]:

$$p(w|a) \approx \sum_{t \in T} p(w|t)p(t|a) \quad a \in A, w \in W.$$

Это приводит к задаче факторизации

$$\Phi^l \approx \Phi^{l+1}\Psi^l,$$

где  $\Phi^{l+1} = [p(w|t)]_{W \times T}$ ,  $\Psi^l = [p(t|a)]_{T \times A}$ . Полученная матрица  $\Psi$  содержит распределения тем  $(l + 1)$ -го уровня в темах  $l$ -го уровня. Таким образом, межуровневый регуляризатор — это логарифм правдоподобия для этой задачи факторизации:

$$R(\Phi, \Psi) = \sum_{a \in A} \sum_{w \in W} n_{wa} \ln \sum_{t \in T} \phi_{wt} \psi_{ta} \rightarrow \max_{\Phi, \Psi},$$

где  $n_a = \sum_{w \in W} n_{wa}$  — веса тем родительского уровня, пропорциональные их размеру.

Этот регуляризатор эквивалентен добавлению в коллекцию  $|A|$  псевдодокументов, представленных матрицей  $[n_{wa}]_{W \times A}$ . Тогда  $\Psi$  образует  $|A|$  дополнительных столбцов матрицы  $\Theta$ , соответствующих этим псевдодокументам.

## 2.4 Метрики качества тематических моделей

Существуют общепринятые в тематическом моделировании методы оценки качества тем. Большинство предложенных методов используют некоторое фиксированное количество  $n$  наиболее вероятных токенов темы (топ-токенов)  $w_i^{(t)}$ , где  $i \in \{1, \dots, n\}$ ,  $t \in T$  и некоторую функцию близости  $f(\cdot, \cdot)$  этих токенов:

$$Q(t) = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n f(w_i^{(t)}, w_j^{(t)}),$$

В [23] предложена мера когерентности темы, основанная на совстречаемости топ-токенов в некоторой коллекции:

$$C(t) = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \ln \frac{D(w_i^{(t)}, w_j^{(t)}) + \varepsilon}{D(w_j^{(t)})} [w_i \neq w_j].$$

Здесь  $D(w_1, w_2)$  – количество документов в некотором корпусе, где слова  $w_1$  и  $w_2$  встретились вместе, а  $D(w)$  – количество документов, в которых встречается токен  $w$ . Для подсчета совстречаемостей предпочтительно использовать большие внешние текстовые коллекции. Когерентность измеряет синтагматическую родственность токенов темы [24].

В работе [25] предложена модификация когерентности, называемая tf-idf когерентностью:

$$C_{tfidf}(t) = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \ln \frac{\sum_{d: w_i \in d, w_j \in d} \text{tfidf}(w_i, d) \text{tfidf}(w_j, d) + \epsilon}{\sum_{d: w_i \in d} \text{tfidf}(w_i, d)} [w_i \neq w_j].$$

Эта метрика помогает решить проблему того, что классическая когерентность слишком сильно полагаться на общие слова, которые часто встречаются в коллекции, но не определяют интерпретируемые темы.

Еще одна метрика такого же типа предложена в [26]. В ней мерой близости токенов является расстояние между их векторными представлениями, то есть векторами модели word embedding:

$$C_{emb}(t) = -\frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n d(v_{w_i}, v_{w_j}) [w_i \neq w_j].$$

Здесь  $v_w$  – вектор, соответствующий токenu  $w$  в пространстве word embedding,  $d(\cdot, \cdot)$  – некоторая функция расстояния между векторами. Метрика, основанная на векторном представлении слов, оценивает парадигматическую родственность токенов темы [24].

Другой тип метрик качества основан на поточечной взаимной информации (PMI). Следуя [27], рассмотрим три метрики этого типа.

В [28] используется расстояние попарного PMI:

$$PMI(t) = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \ln \frac{p(w_i, w_j)}{p(w_i)p(w_j)} [w_i \neq w_j]$$

В [29] используется расстояние нормализованного PMI:

$$NPMI(t) = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \frac{\ln \frac{p(w_i, w_j)}{p(w_i)p(w_j)}}{-\ln p(w_i, w_j)} [w_i \neq w_j].$$

В [23] используется попарная условная вероятность:

$$LCP(t) = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \ln \frac{p(w_i, w_j)}{p(w_i)} [w_i \neq w_j].$$

Главное требование, предъявляемое к метрике качества темы, – согласованность с человеческими оценками. Хорошим качеством согласно метрике должны обладать темы, хорошо интерпретируемые с точки зрения человека, и наоборот. Для проверки того, действительно ли некоторая метрика оценивает интерпретируемость темы, проводят ассессорские эксперименты, в которых собирают мнения людей о качестве некоторого набора тем. Далее проводится сравнение собранных оценок со значениями метрики на темах из эксперимента. Таким образом построены эксперименты, например, в [23] и [26].

В [26] проведено сравнение согласованности с ассессорами всех приведенных метрик качества тем, включая вариации метрики  $C_{emb}(t)$  с использованием различных функций расстояния. Эксперимент проводился на русскоязычных коллекциях, для векторного представления слов использовалась модель word2vec, обученная на большом русскоязычном корпусе из [30]. Самые высокие значения согласованности с ассессорами показала метрика  $C_{emb}(t)$ .

## Глава 3

# Измерение качества тематических иерархий

В обзоре литературы были рассмотрены метрики качества для отдельных тем в тематических моделях. Принятые метрики согласованы с человеческими оценками того, является тема хорошей или плохой.

Тематические иерархии состоят из тем и связей между темами соседних уровней иерархии. Тогда для оценки качества иерархии необходимо измерять не только качество отдельных тем, но и качество отношений "родитель-ребенок" в иерархии. В этой работе мы предлагаем несколько метрик качества для ребер иерархии, которые аппроксимируют мнение ассессоров о наличии или отсутствии связи между темами.

### 3.1 Метрики качества ребер иерархии

#### 3.1.1 Метрики на основе лингвистической близости

Используем такой же вид метрики качества, что используется при оценке качества отдельных тем, для учета синтагматической или парадигматической родственности топ-токенов темы-родителя и темы-ребенка [24]:

$$S(a, t) = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n f(w_i^{(a)}, w_j^{(t)}),$$

Здесь  $w_i^{(s)}$  – это  $i$ -ый топ-токен некоторой темы  $s$ . Тема  $a \in A$  – тема  $l$ -го (родительского) уровня  $t \in T$  – тема  $(l + 1)$ -го (дочернего) уровня,  $f(\cdot, \cdot)$  – мера близости токенов.

Используя ту же меру близости, основанную на совстречаемости

токенов, что используется в классической когерентности из [23], получим оценку сходства между темами

$$S_{coh}(a, t) = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \ln \frac{D(w_i^{(a)}, w_j^{(t)}) + \varepsilon}{D(w_j^{(t)})} [w_i^{(a)} \neq w_j^{(t)}].$$

Здесь  $D(w_1, w_2)$  – количество документов в некотором корпусе, где слова  $w_1$  и  $w_2$  встретились вместе, а  $D(w)$  – количество документов, в которых встречается токен  $w$ .

Используя меру близости из [26], основанную на векторных представлениях слов с косинусным расстоянием между векторами, получим метрику

$$S_{emb}(a, t) = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \langle v(w_i^{(a)}), v(w_j^{(t)}) \rangle [w_i^{(a)} \neq w_j^{(t)}],$$

где  $v(w^{(t)})$  – вектор, соответствующий топ-токену  $w^{(t)}$  темы  $t$  в пространстве word embedding.

### 3.1.2 Метрики на основе вероятностной близости

В тематической иерархии каждая тема представлена вероятностями токенов в ней, значит можно сравнивать родительские и дочерние темы как вероятностные распределения. Используя две стандартные меры расстояния между распределениями – расстояние Хеллингера и дивергенцию Кульбака-Лейблера – получим следующие метрики:

$$S_{Hell}(a, t) = \frac{1}{\sqrt{2}} \|\sqrt{p(w|a)} - \sqrt{p(w|t)}\|_2,$$

$$S_{KL}(a, t) = -D_{KL}(p(w|a) \| p(w|t)).$$

## 3.2 Ассессорская разметка ребер иерархии

Следуя логике, используемой для проверки предлагаемых метрик качества в [23, 26], мы провели ассессорский эксперимент, чтобы собрать человеческие оценки качества ребер иерархии.



### 3.2.1 Описанные данных и моделей

Чтобы собрать пары "родитель-ребенок" для аннотации людьми, мы обучили три двухуровневые иерархические тематические модели на трех коллекциях:

- **Постнаука (Postnauka.ru)**, научно-популярный интернет-журнал с редактируемыми статьями по широкому спектру тем,
- **Хабрахабр (Habrhabr.ru и Geektimes.ru)**, социальные блоги, специализирующиеся на информатике, технологиях и предпринимательстве в сфере IT,
- **Элементы (Elementy.ru)**, научно-популярный веб-сайт с особым упором на естественные науки.

Коллекции состоят из текстовых документов. Коллекции Постнаука и Хабрахабр вручную протэтированы их авторами или редакторами (каждая статья может содержать несколько тегов).

	$ D $	$ W_1 $	$ W_2 $	$ T_1 $	$ T_2 $
ПостНаука	2976	43196	1799	20	58
Хабрахабр	81076	588400	77102	6	15
Элементы	2017	40452	—	9	25

Таблица 3.1: Параметры коллекций.  $|D|$  – размер коллекции,  $|W_1|$  – количество уникальных слов словаря,  $|W_2|$  – количество уникальных тэгов,  $|T_1|$  – количество тем на первом (родительском) уровне,  $|T_2|$  – количество тем на втором (дочернем) уровне.

### 3.2.2 Постановка задания для ассессоров

Эксперимент проводился на краудсорсинг платформе Yandex.Toloka. Участвующим ассессорам был задан следующий вопрос про каждую пару тем: "Даны темы  $T_1$  и  $T_2$ . Является ли одна из тем подтемой другой?". Были следующие варианты ответа: " $T_1$  - это подтема  $T_2$ " " $T_2$  - это подтема  $T_1$ " и "Темы не связаны". Каждая тема  $t$  была обозначена 10 топ-токенами из ее распределения вероятностей  $p(w|t)$ .

После завершения эксперимента первые два варианта ответа были сгруппированы в один ответ «есть связь между темами», поскольку для ассессоров часто было трудно отличить тему-родителя от темы-ребенка по наиболее вероятным словам тем.

### 3.2.3 Контроль качества

Асессоры были отобраны из состава топ-50% экспертов Yandex.Toloka по рейтингу, полученному в ходе всех предыдущих заданий, выполненных экспертом. Перед началом разметки каждый асессор проходил обучение, состоящее из 22 пар тем, которые мы разметили вручную до эксперимента.

Эксперты могли пропустить некоторые задания, если были не уверены в ответе. Асессоры, которые пропустили больше 10 задач подряд отстранялись от участия в эксперименте. Каждому асессору было решено разметить не более 125 ребер.

Каждая пара тем оценивалась пятью разными экспертами.

### 3.2.4 Результаты

В эксперименте приняли участие 68 асессоров, каждый из которых в среднем оценил около 100 пар тем. Оценка одной пары тем в среднем занимала около 5 секунд. В итоге было собрано 6750 оценок 1350 уникальных пар тем.

Участники эксперимента были в основном гражданами России и Украины возрастом от 21 до 64 лет.

Для каждой пары тем мы подсчитали, сколько асессоров дали один и тот же ответ (что темы данной пары связаны или не связаны). В нашем случае, для 5 различных асессоров на каждую пару тем, всегда есть решение большинства о том, связаны ли темы. При этом для каждой темы с ответом большинства могут быть согласны 3, 4 или все 5 асессоров.

Уровень согласия	Количество ребер	Процент ребер
3	374	27.7%
4	468	34.7%
5	508	37.6%

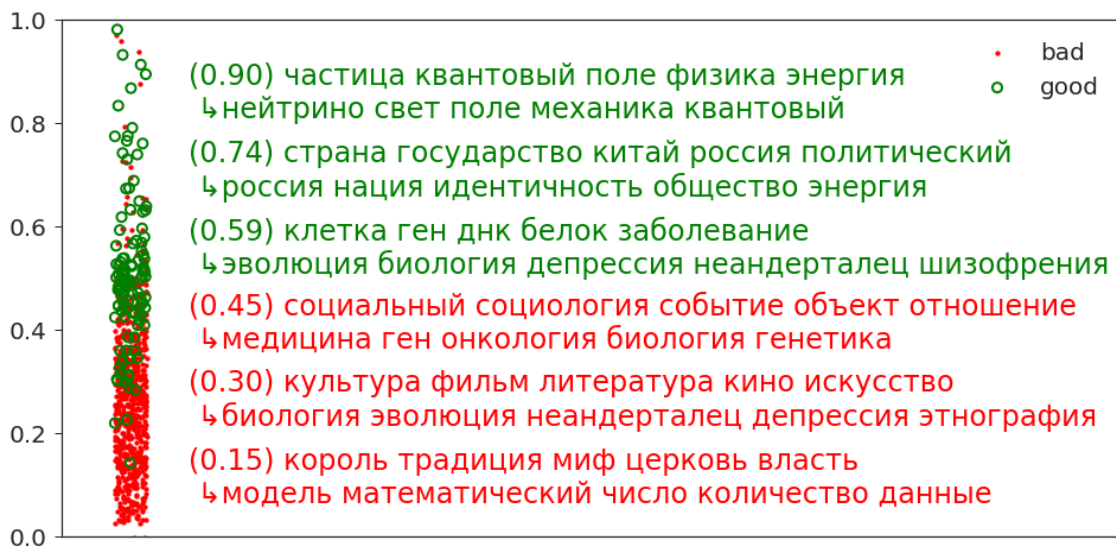
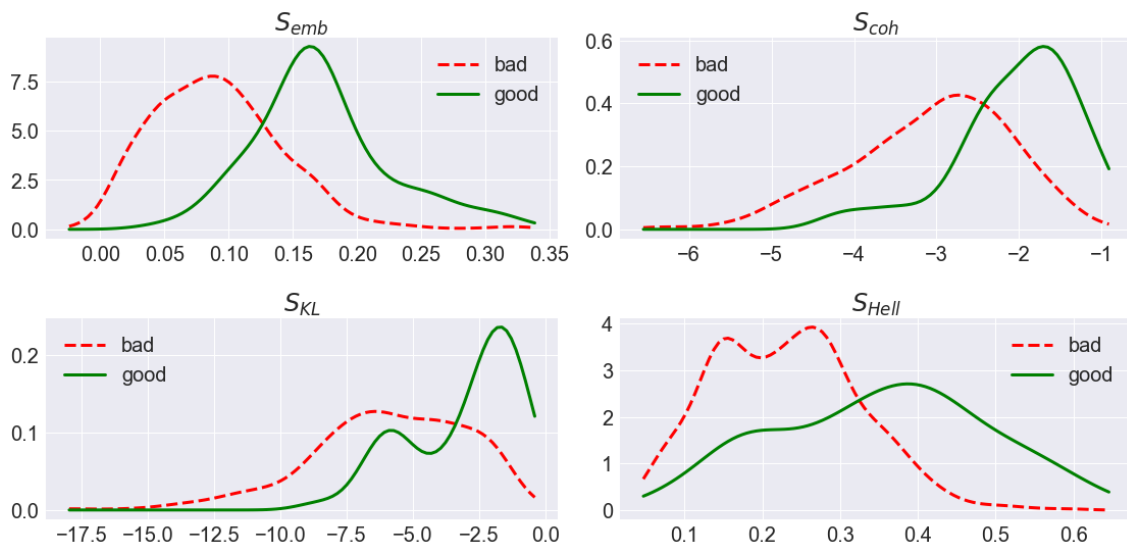
Таблица 3.2: . Согласие асессоров. Уровень согласия – количество асессоров, согласных с решением большинства. Приведены количество и процент от общего числа ребер, соответствующие каждому уровню согласия.

### 3.3 Сравнение метрик с ассессорскими оценками

$$y(T_1, T_2) = \text{sign}(S(T_1, T_2) - w)$$

Метрика	ROC AUC
$S_{emb}$	0.862
$S_{Hell}$	0.766
$S_{KL}$	0.755
$S_{coh}$	0.637

Таблица 3.3: Значения ROC AUC для исследуемых метрик.



### **3.4 Метрики качества тематических иерархий**

## Глава 4

# Агрегирование гетерогенных текстовых коллекций в тематической иерархии

### 4.1 Постановка задачи

### 4.2 Описание данных

Вычислительный эксперимент в данной работе должен проводиться на текстовых коллекциях из нескольких источников, которые существенно отличаются друг от друга по размеру и набору тем. Такие коллекции позволяют исследовать работу предлагаемого метода на гетерогенных данных.

Использовались коллекции статей Хабрахабра и научно-популярного интернет-журнала ПостНаука. Коллекция ПостНауки состоит из 2976 документов и содержит модальности слов (43196 токенов) и тэгов (1799 токенов). Коллекция Хабрахабра состоит из 81076 документов и содержит модальность слов, содержащихся в коллекции ПостНауки (35640 токенов), модальность слов, уникальных для Хабрахабра (545200 токенов), модальность тэгов, содержащихся в коллекции ПостНауки (673 токена), модальность тэгов, уникальных для Хабрахабра (76429 токенов).

Коллекция ПостНауки содержит разнородные темы (технические, естественнонаучные и гуманитарные), поэтому для большинства научно-популярных статей в ней найдутся тематически близкие документы. При этом она сравнительно мала по объему. Коллекция Хабрахабра

напротив содержит в основном технические статьи, меньше естественнонаучных и почти не содержит гуманитарных. При этом объем коллекции большой.

Предобработка данных включает в себя нормализацию текста: перевод в нижний регистр, токенизацию и лемматизацию, удаление редко (менее двух раз) встречающихся слов, удаление наиболее часто встречающихся слов, удаление знаков пунктуации. Кроме того, из коллекций были удалены все статьи, содержащие менее 100 предобработанных слов.

### 4.3 Базовый алгоритм

Стандартный алгоритм построения тематической модели гетерогенных текстовых коллекций предполагает объединение коллекций из разных источников и построение общей модели. В этом эксперименте объединяются коллекции ПостНауки и Хабрахабра.

В силу того, что размер коллекции Хабрахабра существенно превосходит размер коллекции ПостНауки, в большинстве тем полученной модели более 90% статей относятся к Хабрахабру.

Таблица 4.1: Доля статей Хабрахабра в темах модели объединенной коллекции

Номер темы	0	1	2	3	4
Доля статей Хабрахабра	0.941	0.790	0.967	0.983	0.912
Номер темы	5	6	7	8	9
Доля статей Хабрахабра	0.950	0.980	0.994	0.922	0.953
Номер темы	10	11	12	13	14
Доля статей Хабрахабра	0.976	0.905	0.987	0.926	0.977
Номер темы	15	16	17	18	19
Доля статей Хабрахабра	0.996	0.969	0.969	0.991	0.965

Поэтому модель объединенной коллекции отражает в основном тематическую структуру Хабрахабра. Все темы имеют технический характер, характерный для него. При этом потеряны гуманитарные темы, представленные в ПостНауке. Кроме того, объединенная коллекция имеет большой размер, поэтому для построения ее модели требуется намного больше времени, чем для построения модели ПостНауки. Таким образом, базовый алгоритм не решает поставленную задачу и требует модификаций.

## 4.4 Предлагаемый алгоритм

В данной статье предлагается проводить достроение модели в два этапа. На первом этапе проводится фильтрация коллекции, которую необходимо добавить в существующую модель. На втором этапе прошедшие фильтрацию документы добавляются в существующую модель.

### 4.4.1 Фильтрация новой коллекции

Целью этого этапа является сокращение объема добавляемой коллекции (отфильтрованная коллекция должна содержать меньше документов, чем уже содержится в построенной модели) и удаление статей, которые далеки по содержанию от уже присутствующих в модели. Это позволяет отобрать статьи, которые необходимо агрегировать и отсеять те, которые являются нерелевантными. При этом в силу меньшего объема отфильтрованной коллекции возможно сохранить все существующие темы и дополнить те из них, которые наиболее характерны для добавляемой коллекции.

Фильтрацию предлагается проводить по доле слов в статье, присутствующих в словаре существующей модели, и расстоянию до ближайших статей из старой коллекции. Расстояние между документами будем понимать как косинусное расстояние между их tf-idf-представлениями, где idf берется по старой коллекции.

Пусть  $\mathbf{d}_n = [\text{tf}(\text{term}_i, d_n) \cdot \text{idf}(\text{term}_i, D)]_{i=1}^{|W|}$ , где  $D$  — старая коллекция,  $D_{\text{new}}$  — добавляемая коллекция,  $\text{term}_i \in W$  — токены из старого словаря  $W$ ,  $d_n \in D_{\text{new}} \cup D$  — документ.

Расстояние между документами  $d_n$  и  $d_m$  рассчитывается как их косинусное сходство:

$$\text{dist}(d_n, d_m) = \frac{\mathbf{d}_n^T \mathbf{d}_m}{\|\mathbf{d}_n\| \cdot \|\mathbf{d}_m\|}.$$

Доля неуникальных для нового источника слов в документе  $d_n \in D_{\text{new}}$  рассчитывается по формуле

$$P_{\text{common}}(d_n) = \frac{n_{\text{common}}(d_n)}{n_{\text{common}}(d_n) + n_{\text{unique}}(d_n)},$$

где  $n_{\text{common}}$  — количество слов из словаря существующей модели в статье,  $n_{\text{unique}}$  — количество уникальных для источника слов в статье.

Пусть  $\text{Dists} = [\text{dist}(d_n, d_m)]_{d_n \in D_{\text{new}}, d_m \in D}$  — матрица расстояний между документами из новой и старой коллекций. Тогда искомая отфильтрованная выборка состоит из тех документов  $d_n$ , для которых выполняется условие

$$\begin{cases} P_{\text{common}}(d_n) > \text{threshold}_1 \\ \text{mean}_{m \in [m_1, \dots, m_{10}]}(\text{Dists}[n, m]) < \text{threshold}_2, \end{cases}$$

где  $[m_1, \dots, m_{10}]$  — индексы, соответствующие 10 наименьшим значениям  $\text{Dists}[n, :]$ ,  $\text{mean}(\cdot)$  — среднее арифметическое,  $\text{threshold}_i, i \in [1, 2]$  — заданные пороги.

#### 4.4.2 Дополнение модели отфильтрованной коллекцией

Полученная выборка документов добавляется в существующую модель. Для этого объединим старую коллекцию с этой выборкой.

**Первый уровень иерархии:** Предполагается, что первый уровень иерархической тематической модели содержит все темы, характерные для агрегируемого контента. Тогда после добавления новых документов строки новой матрицы  $\Phi_{\text{new}}^1$  первого уровня иерархии, соответствующие токенам старого словаря, не должны значительно отличаться от соответствующих строк старой матрицы  $\Phi^1$ . Поэтому инициализируем эту подматрицу  $\Phi_{\text{new}}^1$  матрицей  $\Phi^1$ , а остальные строки инициализируем случайно. Для того чтобы темы не изменились значительно, применим регуляризатор сглаживания  $\Phi^1$  по всем инициализированным темам.

**Второй уровень иерархии:** Предполагается, что второй уровень иерархии содержит более специфические темы, чем первый уровень. Тогда добавление новых документов в коллекцию может привести к появлению подтем, характерных для нового источника, на втором уровне. Поэтому добавим в модель некоторое фиксированное количество новых тем. Инициализируем подматрицу  $\Phi_{\text{new}}^2$ , соответствующую старым токенам и темам, старой матрицей  $\Phi^2$ , как и на первом уровне. Новые токены и темы инициализируем случайно. Для сохранения инициализированных тем применяем регуляризатор сглаживания  $\Phi$  по ним.

Новые темы должны быть специфичными для нового источника, то есть большинство отнесенных к ним документов должны принадлежать новой коллекции. Поэтому применим регуляризатор разреживания  $\Theta$  по новым темам для документов старой коллекции.

### 4.5 Сравнение алгоритмов



## **Глава 5**

# **Заключение**

# Литература

- [1] David M. Blei, “Probabilistic topic models”, *Commun. ACM*, vol. 55, no. 4, pp. 77–84, 2012.
- [2] T. Hoffman, “Probabilistic latent semantic indexing”, in *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 50–57. ACM Press, New York, 1999.
- [3] David M. Blei, Andrew Y. Ng, and Michael I. Jordan, “Latent dirichlet allocation”, *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.
- [4] C Chemudugunta, P Smyth, and M Steyvers, “Modeling general and specific aspects of documents with a probabilistic topic model”, *Nips*, 2006.
- [5] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth, “The author-topic model for authors and documents”, *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, 2004.
- [6] Khoat Than and Tu Bao Ho, “Fully sparse topic models”, in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2012.
- [7] Konstantin Vorontsov, Oleksandr Frei, Murat Apishev, Peter Romov, Marina Suvorova, and Anastasia Yanina, “Non-bayesian additive regularization for multimodal topic modeling of large collections”, in *TM@CIKM*, Nikolaos Aletras, Jey Han Lau, Timothy Baldwin, and Mark Stevenson, Eds. 2015, pp. 29–37, ACM.
- [8] Konstantin Vorontsov, Anna Potapenko, and Alexander Plavin, “Additive regularization of topic models for topic selection and sparse factorization”, *Statistical Learning and Data Sciences*, January 2015.

- [9] K. V. Vorontsov, “Additive regularization for topic models of text collections”, *Doklady Mathematics*, vol. 89, no. 3, pp. 301, May 2014.
- [10] Konstantin Vorontsov and Anna Potapenko, “Additive regularization of topic models”, *Machine Learning*, vol. 101, no. 1-3, pp. 303, October 2015.
- [11] Konstantin Vorontsov and Anna Potapenko, “Tutorial on probabilistic topic modeling: Additive regularization for stochastic matrix factorization”, in *Analysis of Images, Social Networks and Texts*. Springer, January 2014.
- [12] D. M. Blei, T. Griffiths, Michael I. Jordan, and J. Tenenbaum, “Hierarchical topic models and the nested chinese restaurant process”, 2003.
- [13] David M. Mimno, Wei Li 0010, and Andrew McCallum, “Mixtures of hierarchical topics with pachinko allocation”, in *ICML*, Zoubin Ghahramani, Ed. 2007, vol. 227 of *ACM International Conference Proceeding Series*, pp. 633–640, ACM.
- [14] NA Chirkova and KV Vorontsov, “Additive regularization for hierarchical multimodal topic modeling”, 2016.
- [15] Chi Wang, Jialu Liu, Nihit Desai, Marina Danilevsky, and Jiawei Han, “Constructing topical hierarchies in heterogeneous information networks”, *Knowl. Inf. Syst.*, vol. 44, no. 3, pp. 529–558, 2015.
- [16] Marina Danilevsky, Chi Wang, Fangbo Tao, Son Nguyen, Gong Chen, Nihit Desai, Lidan Wang, and Jiawei Han, “AMETHYST: A system for mining and exploring topical hierarchies of heterogeneous data”, December 03 2013.
- [17] Nachiketa Sahoo, Jamie Callan, Ramayya Krishnan, George T. Duncan, and Rema Padman, “Incremental hierarchical clustering of text documents”, in *CIKM*, Philip S. Yu, Vassilis J. Tsotras, Edward A. Fox, and Bing Liu 0001, Eds. 2006, pp. 357–366, ACM.
- [18] 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*. Springer, January 2015.

- [19] Oleksandr Frei and Murat Apishev, “Parallel non-blocking deterministic algorithm for online topic modeling”, in *Analysis of Images, Social Networks and Texts*. Springer, January 2017.
- [20] Elias Zavitsanos, Paliourg@iit Demokritos Gr, George A Vouros, and Georgev@aegean Gr, “Non-Parametric Estimation of Topic Hierarchies from Texts with Hierarchical Dirichlet Processes Georgios Paliouras”, *Journal of Machine Learning Research*, 2011.
- [21] Jay Pujara and Peter Skomoroch, “Large-Scale Hierarchical Topic Models”, *NIPS Workshop on Big Learning*, 2012.
- [22] Ke Zhai, Jordan Boyd-Graber, Nima Asadi, and Mohamad L. Alkhouja, “Mr. LDA”, in *Proceedings of the 21st international conference on World Wide Web - WWW '12*, New York, New York, USA, 2012, p. 879, ACM Press.
- [23] David Mimno, Hanna M. Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum, “Optimizing semantic coherence in topic models”, in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 2011.
- [24] Hinrich Schütze and Jan Pedersen, “A Vector Model for Syntagmatic and Paradigmatic Relatedness”, *Making Sense of Words: Proceedings of the Conference*, 1993.
- [25] Sergey I. Nikolenko, Sergei Koltcov, and Olessia Koltsova, “Topic modelling for qualitative studies”, *Journal of Information Science*, vol. 43, no. 1, pp. 88–102, feb 2017.
- [26] Sergey I. Nikolenko and Sergey I., “Topic Quality Metrics Based on Distributed Word Representations”, in *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval - SIGIR '16*, 2016.
- [27] Jey Han Lau, David Newman, and Timothy Baldwin, “Machine Reading Tea Leaves: Automatically Evaluating Topic Coherence and Topic Model Quality”, in *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, Stroudsburg, PA, USA, 2014, pp. 530–539, Association for Computational Linguistics.
- [28] David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin, “Automatic evaluation of topic coherence”, in *Human Language*

*Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics.* Association for Computational Linguistics, 2010, pp. 100–108.

- [29] Gerlof Bouma, “Normalized (pointwise) mutual information in collocation extraction”, *Proceedings of GSCL*, pp. 31–40, 2009.
- [30] A. Panchenko, N. V. Loukachevitch, D. Ustalov, D. Paperno, C. M. Meyer, and N. Konstantinova, “Russe: the First Workshop on Russian Semantic Similarity”, *Компьютерная Лингвистика И Интеллектуальные Технологии: По Материалам Ежегодной Международной Конференции «Диалог»*, 2015.