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 М. С.

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

Москва, 2018

Оглавление

Глава 1

Введение

Задача агрегирования и классификации знаний для поиска литературы существует с момента появления больших библиотек. Используемые в библиотеках методы каталогизации и классификации книг [1] существенно опираются на ручной труд и требуют привлечения экспертов для расширения или изменения структуры каталогов.

С другой стороны, обработка естественного языка позволяет автоматизировать решение многих задач работы с текстами. Так тематическое моделирование с момента своего создания успешно применялось для навигации по крупным текстовым корпусам и их визуализации [2, 3, 4]. В больших текстовых коллекциях темы часто образуют иерархии, в которых каждая тема делится на более специфичные подтемы. Моделью такой ситуации являются иерархические тематические модели. Они удобны для навигации по коллекциям, поэтому являются подходящей моделью для агрегирования контента.

В этой работе рассматривается задача построения иерархических тематических моделей на гетерогенных данных, собранных из различных источников. На сегодня нет общепринятых методов построения интерпретируемых тематических иерархий. Одна из проблем состоит в том, что для них не существует стандартных метрик качества, что делает невозможным сравнение различных моделей друг с другом. В работе предложены метрики качества тематических иерархий, согласованные с мнением людей об интерпретируемости иерархий. Для сравнения метрик с оценками людей проведен асессорский эксперимент. Кроме того, предложен метод построения иерархических тематических моделей гетерогенных данных для задачи агрегирования. Проведено сравнение нескольких алгоритмов и показано, что предложенный метод дает наилучшее качество и существенно превосходит базовый подход.

Вероятностное тематическое моделирование — это раздел машин-

ного обучения, решающий задачу поиска тем в коллекции документов. Тематическая модель определяет к каким темам относится каждый документ и какие слова образуют темы [5]. Базовыми подходами построения тематических моделей являются PLSA [6] и LDA [7]. В работах [8, 9, 10] предложены модификаций данных подходов, учитывающие специфику конкретных задач. Аддитивная регуляризация тематических моделей (ARTM) [11, 12, 13, 14] позволяет комбинировать различные модели, интерпретируя их как регуляризаторы в PLSA.

Общепринятого определения и подхода к построению иерархических тематических моделей не существует. В модели иерархического LDA (hLDA) [15] темы образуют дерево. С другой стороны, модель иерархического распределения патинко (hPAM) [16] и модель иерархического ARTM (hARTM) [17] представляют собой направленный ациклический многодольный граф, что лучше соответсвует реальным отношениям между темами в мультидисциплинарных статьях.

В данной работе используется модель hARTM [17]. Это развитие идеи аддитивной регуляризации тематических моделей для задачи построения тематических иерархий. Она позволяет применять регуляризацию как к темам всех уровней иерархии для комбинирования любых тематических моделей, так и к самой иерархии для контроля разреженности отношения «родитель-ребенок».

В работах [18, 19, 20, 21, 22, 23] предложены различные метрики качества для отдельных тем в тематической модели. Большинство метрик используют различные меры близости наиболее вероятных слов темы для оценки ее качества. В таком случае качество всей модели – это некоторая функция от качества ее тем, например, их среднее качество.

Тематическая иерархия состоит из тем и ребер иерархии, характеризующих связи между темами. Так как уже существуют принятые метрики качества тем, в данной работе для решения задачи оценки качества иерархии предлагается ввести метрики качества также для ребер иерархии. Предлагаемые метрики основаны на близости между наиболее вероятными словами темы-ребенка и темы-родителя. В асессорском эксперименте собраны мнения людей на тему наличия или отсутствия связи между темами. Показано, что предлагаемые метрики хорошо аппроксимируют человеческие оценки того, связаны темы некоторой пары или нет.

Базовый подход к построению общей тематической модели гетерогенных коллекций, собранной из нескольких источников, различных по объему и тематической структуре, предполагает слияние коллекций в одну и построение ее модели. При таком подходе темы, уникаль-

ные для меньшего из источников, теряются. Кроме того, для задачи агрегирования важно автоматически выбирать контент, который нужно включать в модель и отфильтровывать документы, не относящиеся к агрегируемому контенту. В данной работе для решения задачи агрегирования предлагается дополнять существующую качественную модель одного источника (базовой коллекции), построенную заранее, выборками документов из новых источников. Количество добавленных документов может во много раз превышать размер базовой коллекции. При этом документы, которые добавляются в модель, выбираются на основе близости к базовой коллекции. При обучении модели используется инициализации параметров модели дополненной коллекции параметрами модели базовой коллекции и сокращение словаря модели до словаря базовой коллекции. Такой подход позволяет сохранить качество модели и распределить новые документы по подходящим темам, сохранив при этом темы базовой коллекции.

В вычислительных экспериментах в данной работе используются коллекции русскоязычного научно-популярного контента. Для проведения экспериментов используется BigARTM — библиотека для тематического моделирования с открытым исходным кодом [24, 25].

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 $n_d = \sum\limits_{w \in W} n_{dw}$ — 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

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

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) [6]. В ней матрицы Φ и Θ находятся с помощью максимизации логарифма правдоподобия:

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

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 EM-алгоритмом:
E-шаг:

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

M-шаг:

$$\phi_{wt} = \frac{n_{wt}}{\sum\limits_{v \in W} n_{vt}}; \quad n_{wt} = \sum_{d \in D} n_{dw} p_{tdw};$$

$$\theta_{td} = \frac{n_{td}}{\sum\limits_{s \in T} n_{sd}}; \quad n_{td} = \sum_{w \in d} n_{dw} p_{tdw}.$$

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) [7], в которой вектора ϕ_t и θ_d генерируются из распределений Дирихле. Кроме того, было предложено множество обобщений LDA, учитывающих дополнительные требования к модели [26, 27, 28].

2.2.3 Аддитивная регуляризация тематических моделей

Априорным распределениям параметров тематической модели соответствуют регуляризаторы в задаче оптимизации. Аддитивная регуляризация тематических моделей (модель ARTM) [14] позволяет одновременно учитывать множество дополнительных требований к модели.

В ARTM матрицы Φ и Θ находятся с помощью максимизации суммы логарифма правдоподобия и регуляризаторов R_i с неотрицательными коэффициентами регуляризации τ_i:

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

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

Как показано в [14], EM-алгоритм для этой задачи имеет следующий вид:

E-шаг:

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

M-шаг:

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

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

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

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 ϕ_t и θ_d были близки к заданным распределениям $\beta_t = [\beta_{wt}]_{w\in W}$ и $\alpha_d = [\alpha_{td}]_{t\in T}$ в смысле дивиргенции Кульбака-Лейблера [14]:

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

где β_0 и α_0 — заданные положительные коэффициенты. Введение этого регуляризатора, как показано в [14], эквивалентно модели LDA.

Регуляризатор разреживания имеет такой же вид, но коэффициенты β_0 и α_0 отрицательны. Он способствует обращению значительной части вероятностей ϕ_{wt} и θ_{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 ϕ_t, что способствует увеличению различности тем модели [14]:

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

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 $m \in M$ соответсвует отдельный словарь (множество токенов) W^m, причем $W = \bigsqcup\limits_{m\in M} W^m$.

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

строк F^m и Φ^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 [14]:

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

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

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) [15]. В нем темы образуют дерево, то есть каждая подтема (тема-ребенок) имеет только одну тему-родителя.

С другой стороны, модель иерархического распределения патинко (hPAM) [16] представляет собой направленный ациклический многодольный граф. Модель иерархического процесса Дирихле (hHDP) [29] 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 [30, 31].

2.3.2 Иерархическая модель ARTM

В иерархической модели ARTM (hARTM) [17] 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 Φ, Θ и множеством тем T.

Будем моделировать распределение токенов по темам l-того уровня как смесь распределений по темам $(l+1)$-го уровня [17]:

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

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 Ψ содержит распределения тем $(l+1)$-го уровня в темах l-го уровня. Таким образом, межуровневый регуляризатор — это логарифм правдоподобия для этой задачи факторизации:

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

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

Этот регуляризатор эквивалентен добавлению в коллекцию $|A|$ псевдодокументов, представленных матрицей $[n_{wa}]_{W \times A}$. Тогда Ψ образует $|A|$ дополнительных столбцов матрицы Θ, соответсвующих 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 $i \in \{1, ..., n\},\ t \in T$ и некоторую функцию близости $f(\cdot, \cdot)$ этих токенов:

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

В [18] 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 [32].

В работе [20] предложена модификация когерентности, называемая tf-idf когерентностью:

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

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 [19]. 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 токену w в пространстве word embedding, $d(\cdot,\cdot)$ – некоторая функция расстояния между векторами. Метрика, основанная на векторном представлении слов, оценивает парадигматическую родственность токенов темы [32].

Другой тип метрик качества основан на поточечной взаимной информации (PMI). Следуя [21], рассмотрим три метрики этого типа.

В [22] 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 [23] 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 [18] 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 асессорские 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 [18] и [19].

В [19] проведено сравнение согласованности с асессорами 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 [33]. Самые высокие значения согласованности с асессорами 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 предлагается несколько метрик качества для ребер иерархии, которые аппроксимируют мнение асессоров 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 [32]:

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

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 [18], получим оценку сходства между темами

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

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 [19], основанную на векторных представлениях слов с косинусным расстоянием между векторами, получим метрику

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

где $v(w^{(t)})$ – вектор, соотвествующий 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:

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

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

3.2 Асессорская разметка ребер иерархии

Следуя логике, используемой для проверки предлагаемых метрик качества в [18, 19], был проведен асессорский эксперимент, чтобы собрать человеческие оценки качества ребер иерархии.

3.2.1 Описанные данных и моделей

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

- **Постнаука (Postnauka.ru)**, научно-популярный интернет-журнал с редактируемыми статьями по широкому спектру тем,
- **Хабрахабр (Habrahabr.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 протэгированы 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 задачи для асессоров

Эксперимент проводился на краудсорсинговой платформе Yandex.Toloka. Участвующим асессорам 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 асессоров 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 были размечены 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 разрешено 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 было подсчитано, 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 асессорскими оценками

Для того, чтобы показать согласованность метрик с оценками асессоров, введем следующую задачу классификации. Пары тем из асессорского эксперимента делятся на два класса: «хорошие» и «плохие». «Хорошие» пары – те, для которых не менее 4 асессоров согласились с утверждением о том, что темы связаны. Если пара (T_1, T_2) «хорошая» , то в задаче классификации тем верный ответ $y(T_1, T_2) = 1$, иначе тема «плохая» и $y(T_1, T_2) = -1$. Тогда каждая метрика задает классификатор:

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

где T_1 и T_2 – пара тем из родительского и дочернего уровней иерархии соответственно, S – одна из предложенных метрик качества, а w – отступ классификатора.

Поставив задачу таким образом, можно рассчитать ROC AUC классификаторов и оценить таким образом качество каждой из метрик: большие значения ROC AUC соответствуют лучшей аппроксимации асессорских оценок.

Метрика	ROC AUC
S_{emb}	0.878
S_{Hell}	0.815
S_{KL}	0.790
S_{coh}	0.766

Таблица 3.3: Значения ROC AUC для исследуемых метрик.

В таблице 3.3 указаны значения ROC AUC для всех классификаторов. Наилучшее качество показал классификатор, основанный на метрике S_{emb} (AUC = 0.878). Остальные метрики также показали приемлемое качество классификации: все значения AUC выше 0.75. Для лучшего понимания результата, приведем график 3.1. Для каждой метрики на соответствующем графике красная (пунктирная) линия соответствует плотности вероятности значений этой метрики (по оси x) для «плохих» ребер, а зеленая (сплошная) линия соответсвует плотности вероятности значений для «хороших» ребер. Чем сильнее разнесены эти распределения, тем лучше значения метрики согласованы с асессорскими оценками.

В дальнейшем в вычислительных экспериментах используется метрика S_{emb}, так как она показала наилучшую согласованность.

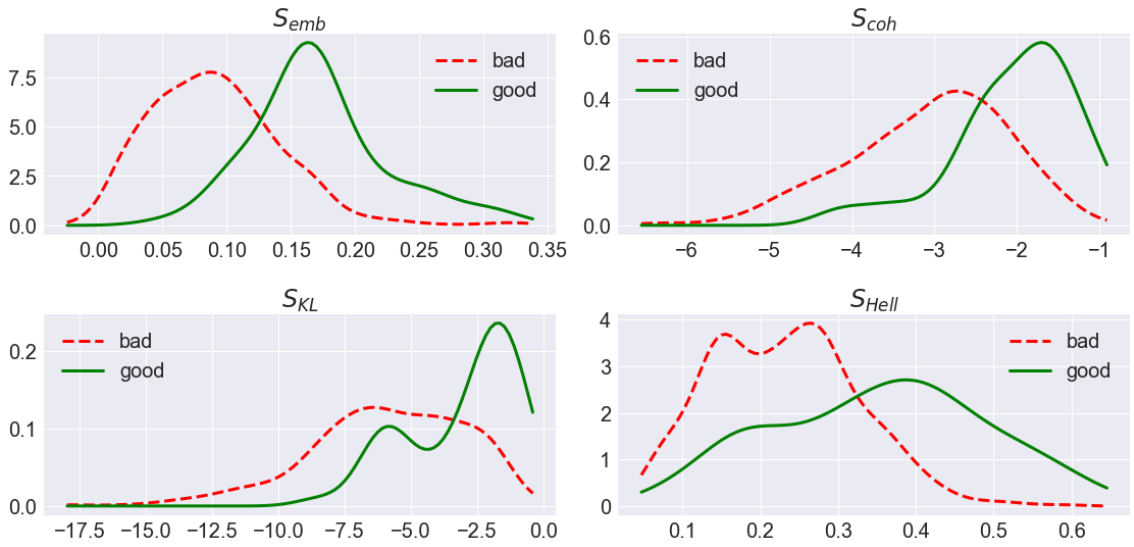

Рис. 3.1: Вероятностные распределения значений метрик для «хороших» и «плохих» ребер.

Приведем также качественную иллюстрацию работы метрик. На рисунке 3.2 приведены 6 пар тем, которые оценивали асессоры в эксперименте. Три из них были оценены как «хорошие», три оставшиеся были оценены как «плохие».

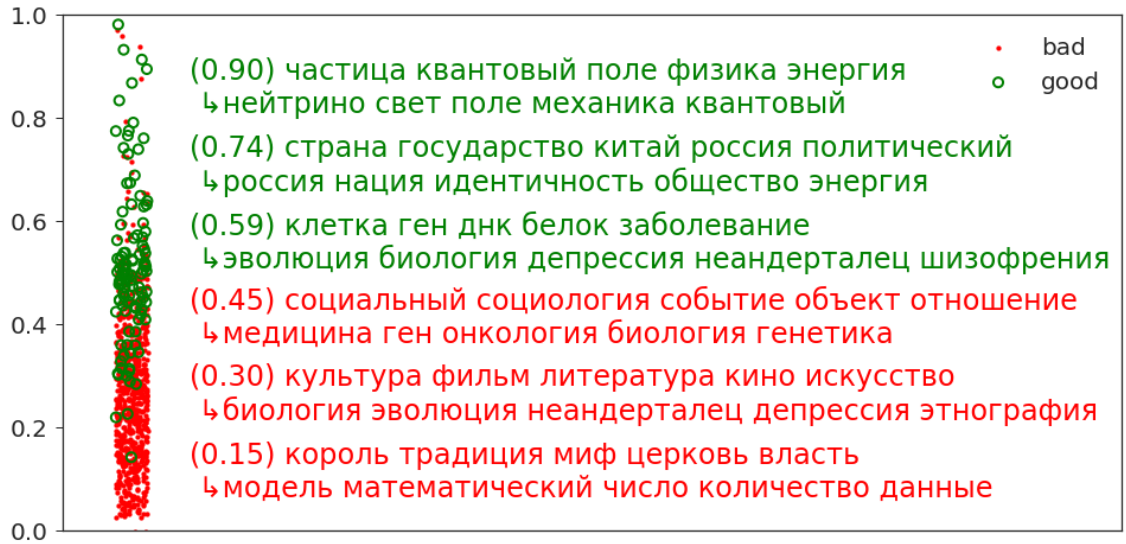

Рис. 3.2: Примеры пар тем из асессорского эксперимента с соответствующими им значениями метрики S_{emb} и вердиктом асессоров. Каждая тема и подтема представлены своими 5 топ-токенами.

Для каждой пары указано соответствующее ей значение метрики S_{emb}. Слева находится распределение всех ребер из асессорского эксперимента. Y-координаты точек соответствуют значениям метрики S_{emb} для пар тем, а цвета показывают мнение асессоров. Видно, что в экс-

перименте было гораздо больше пар, оцененных как «плохие», чем тех, которые оценены как «хорошие». Поскольку ожидается, что тематическая иерархия является разреженной (каждая родительская тема имеет только небольшое количество подходящих подтем), это наблюдение соответствует нашим ожиданиям. Из рисунка также ясно, что «плохие» пары имеют более низкое среднее значение метрики, чем «хорошие». Это означает, что метрика S_{emb} оценивает «хорошие» пары более высокими значениями и, следовательно, коррелирует с мнением асессоров.

3.4 Метрики качества тематических иерархий

Цель состоит в том, чтобы объединить оценки качества ребер в некоторую конструкцию, которая была бы представительной мерой качества для иерархии в целом. Предлагается несколько подходов к этой задаче. Тогда, оценив общее качество связей иерархии с помощью наших метрик и общее качество тем иерархии с помощью метрик для плоских моделей, можно получить представление о качестве иерархии и использовать полученные оценки для сравнения разных моделей.

3.4.1 Среднее качество ребер

В [18] качество плоской тематической модели оценивается как среднее качество тем этой модели. Следуя этому же принципу, оценим качество связей иерархии как среднее значение качества ребер.

Как описано в секции 2.3.2, связи в тематической иерархии задаются вероятностями $p(t|a)$ для пар тем из соседних уровней быть родителем и ребенком. Таким образом, конкретная конфигурация иерархии зависит от порога вероятности, которая является достаточной для включения ребра, соединяющего t и a, в иерархию. Поэтому разные пороговые значения приводят к различным значениям среднего качества ребер. Для разных моделей значения $p(t|a)$ могут лежать в разных интервалах. Для того, чтобы сравнивать разные иерархические модели при одинаковых значениях порога, будем здесь и далее использовать нормализованную матрицу Ψ:

$$\psi_{ta}^{norm} = \frac{\psi_{ta} - \min_t \psi_{ta}}{\max_t \psi_{ta} - \min_t \psi_{ta}}.$$

Тогда $\Psi^{norm} = [\psi_{ta}^{norm}]_{T \times A}$ и для любой модели вероятности имеют значения от 0 до 1.

Недостатком предложенной метрики качества связей иерархии является неинтерпретируемость ее абсолютного значения.

3.4.2 Качество ранжирования

Второй предлагаемый метод оценки качества связей иерархии рассматривает качество ранжирования ребер по значениям вероятностей в матрице Ψ. Предположим, что у нас есть несколько уровней иерархии, представляющих собой плоские тематические модели. Если нужно провести фиксированное количество k ребер между темами этих уровней, то естественно выбрать ребра между наиболее связанными с точки зрения человека парами тем. Как показано ранее в секции 3.3, аппроксимацией такого выбора будет набор k ребер с наибольшими значениями метрики S_{emb}. Таким образом, за правильный ответ ранжирования можно взять ранжирование, которое дает метрика и оценить качество ранжирования, которое задает Ψ с помощью принятых метрик качества ранжирования, таких как:

- Средняя точность (Average Precision, AP@k) – описана в [34].
- Inverse Defect Pairs, IDP@k – обратное к значению количества пар, которые ранжируются алгоритмом в неправильном порядке.
- Normalized Discounted Cumulative Gain, nDCG@k – описана в [34].

Глава 4

Агрегирование гетерогенных текстовых коллекций в тематической иерархии

4.1 Постановка задачи

В данной работе решается задача агрегирования больших гетерогенных текстовых коллекций в их общей иерархической тематической модели. Предполагается, что есть начальное приближение тематической иерархии, в котором уже присутствует большинство агрегируемых тем.

Назовем коллекцию, модель которой берется за начальное приближение, базовой. Такая модель должна быть хорошо интерпретируемой. Алгоритм, решающий поставленную задачу, должен позволить увеличивать объем и разнообразие агрегируемого контента, добавляя в модель базовой коллекции документы из других источников. При этом конечный размер агрегируемой коллекции может во много раз превышать размер базовой коллекции. Кроме того, модель должна оставаться интерпретируемой и как можно меньше терять в качестве по сравнению с моделью базовой коллекции.

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.

В качестве примера в этой работе рассматривается задача агреги-

рования русскоязычного научно-популярного контента из коллекций, описанных в секции 3.2.1. Будем использовать коллекцию ПостНауки в качестве базовой, так как она 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 ее объем во много раз превышает объем коллекции ПостНауки, взятой за базовую. Таким образом, на примере этих коллекций можно оценить эффективность алгоритмов для задачи агрегации.

4.2 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.

t	0	1	2	3	4	5	6	7	8	9	10
P_t	0.94	0.79	0.97	0.98	0.91	0.95	0.98	0.99	0.92	0.95	0.98
t	10	11	12	13	14	15	16	17	18	19	20
P_t	0.98	0.91	0.99	0.93	0.98	1.0	0.97	0.97	0.99	0.97	0.95

Таблица 4.1: Доля статей Хабрахабра в темах модели объединенной коллекции. t – номер темы, P_t – доля статей Хабрахабра в этой теме, если к теме относятся документы с $p(t|d) > 0.1$.

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.

4.3 Предлагаемый алгоритм

В данной работе для решения поставленной задачи предлагается проводить построение модели в несколько этапов. Обозначим базовую коллекцию как D_0, а коллекцию, состоящую из документов, которые необходимо добавить к агрегируемому контенту, как D_1. Тогда предлагаемый метод можно описать следующим образом:

1. **Начальное приближение**: построить модель базовой коллекции D_0. Параметры модели начального приближения обозначим как Φ_0^l, Θ_0^l, где l – номер уровня.

2. **Фильтрация**: отранжировать документы добавляемой коллекции D_1 в порядке уменьшения близости к базовой коллекции D_0 в смысле некоторой метрики.

3. **Дополнение модели:** добавить выбранное количество документов D_1, отранжированных в начало списка на этапе фильтрации, в коллекцию. При этом получим новую коллекцию D. Инициализировать параметры Φ^l модели коллекции D параметрами Φ_0^l начального приближения. Обучить иерархическую модель D.

Первый этап – задача построения интерпретируемой тематической иерархии однородной коллекции. Эта задача рассмотрена в [17], кроме того, в [35] рассмотрена задача подбора коэффициентов регуляризации в модели ARTM.

Опишем подробнее два последних этапа.

4.3.1 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 присутсвующих в модели. Это позволяет отобрать статьи, которые необходимо агрегировать и отсеять те, которые являются нерелевантными.

Фильтрацию предлагается проводить по расстоянию новых документов до ближайших документов из старой коллекции. Это расстояние можно рассчитывать по разным метрикам близости документов. В этой работе будем использовать косинусное расстояние между их tf-idf-представлениями, где idf берется по старой коллекции.

Пусть $\boldsymbol{d_n} = [\text{tf}(w_i, d_n) \cdot \text{idf}(w_i, D_0)]_{i=1}^{|W|}$, где D_0 — базовая коллекция, D_1 — добавляемая коллекция, $w_i \in W$ — токены из словаря базовой коллекции W, $d_n \in D_0 \cup D_1$ — документ.

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

$$\rho(d_n, d_m) = \frac{\boldsymbol{d_n}^T \boldsymbol{d_m}}{||\boldsymbol{d_n}|| \cdot ||\boldsymbol{d_m}||}.$$

Для каждого документа $d_n \in D_1$ можно найти 10 ближайших документов из коллекции D_0. Пусть этим документам соответствуют индексы $[m_1, ..., m_{10}]$. Тогда искомая отфильтрованная выборка состоит из тех документов d_n, для которых выполняется условие

$$\frac{1}{10} \sum_{m \in [m_1, ..., m_{10}]} \rho(d_n, d_m) < t,$$

где t – некоторое заданное пороговое сходство между документами.

Обозначим отфильтрованную коллекцию, состоящую из документов, удовлетворяющих данному условию, как D_f.

4.3.2 Дополнение модели отфильтрованной коллекцией

На этом этапе полученная выборка документов добавляется в модель начального приближения иерархии. При этом предполагается, что начальное приближение уже содержит все темы, характерные для агрегируемого контента. Тогда процесс добавления новых документов можно описать следующим образом:

1. **Слияние коллекций:** добавить отфильтрованную коллекцию к базовой и получить объединенную коллекцию $D = D_0 \cup D_f$.

2. **Инициализация:** после добавления новых документов строки новых матриц Φ^l иерархии коллекции D, соответсвующие токенам старого словаря, не должны значительно отличаться от соответствующих строк старой матрицы Φ_0^l. Будем строить модель, используя только словарь базовой коллекции. Поэтому инициализируем матрицы Φ^l соответствующими матрицами Φ_0^l.

3. **Обучение иерархии:** используем модель hARTM для построения иерархической тематической модели коллекции D. На каждом уровне иерархии начальные значения параметров заданы на этапе инициализации.

4.4 Вычислительный эксперимент: агрегирование русскоязычного научно-популярного контента

Опишем реализацию предложенного метода для задачи агрегирования русскоязычного научно-популярного контента на примере коллекций ПостНауки и Хабрахабра.

4.4.1 Начальное приближение: иерархия ПостНауки

Первый этап — построение модели ПостНауки, в которую в дальнейшем планируется добавлять документы Хабрахабра.

Модель содержит два уровня иерархии. На первом уровне 21 тема (одна фоновая). На втором уровне 61 тема (1 фоновая). Использовались модальности слов и тэгов, причем для тэгов был установлен значительно больший вес, что позволило существенно улучшить качество модели.

Первый уровень иерархии

Применялись регуляризаторы декорреляции терминов нефоновых тем и сглаживания Ф для фоновой темы.

Рис. 4.1: Визуализация первого уровня иерархии ПостНауки

На рисунке 4.1 изображена визуализация полученной модели. В названия тем вынесены 3 наиболее вероятных для данной темы токена.

Второй уровень иерархии

Применялась следующая стратегия регуляризации:

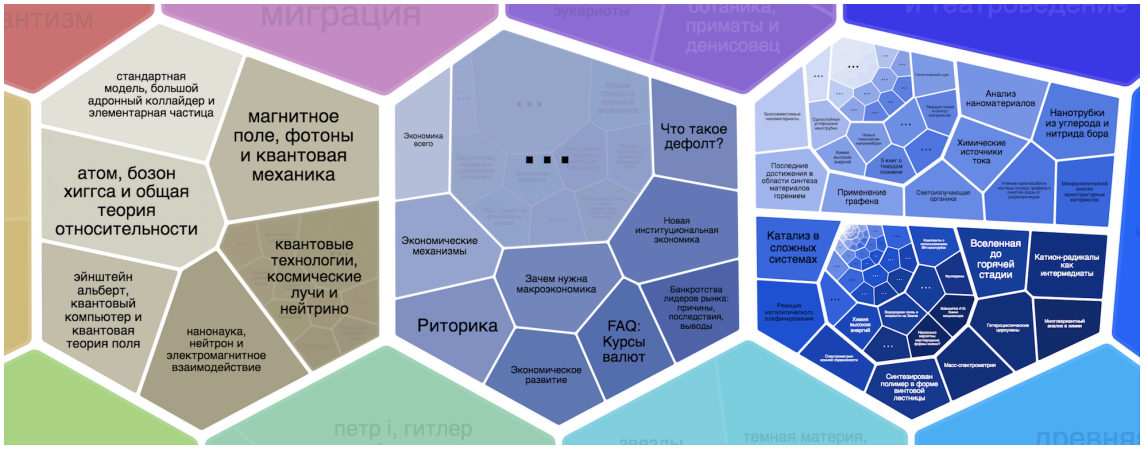

Рис. 4.2: Визуализация уровней иерархии ПостНауки. Первый и второй уровень содержат темы, на третьем уровне находятся отнесенные к темам документы.

- первые 10 итераций применялись регуляризаторы декорреляции терминов нефоновых тем и сглаживания Φ фоновой темы;
- после 10 итераций подмножества нефоновых тем, которые модель отнесла к одной родительской теме первого уровня, разреживались с помощью регуляризатора разреживания Θ между собой с коэффициентами τ, возрастающими с каждой итерацией и пропорциональными количеству тем в каждом подмножестве.

На рисунке 4.2 изображены различные уровни иерархии в полученной модели, темы проименованные тремя наиболее вероятными токенами.

Полученная модель содержит существенно различные темы в силу того, что контент ПостНауки разнородный (содержит как естественнонаучные, так и гуманитарные и технические статьи). В дальнейшем будем предполагать, что в ней содержатся все характерные для научно-популярных статей темы.

4.4.2 Фильтрация коллекции Хабрахабра

Коллекция Хабрахабра была отранжирована по предложенной метрике близости. Затем полученные значения расстояний были отнормированы по метрике max. Было выбрано значение порогового расстояния $t = 0.85$. При этом в отфильтрованную коллекцию D_f вошло 9288 документов, то есть размер добавленной коллекции в несколько раз превышает размер базовой коллекции.

4.4.3 Дополнение модели ПостНауки отфильтрованной коллекцией Хабрахабра

Модель содержит два уровня иерархии. На первом уровне 21 тема (одна фоновая). На втором уровне 61 тема (1 фоновая). Объединенная коллекция $D = D_0 \cup D_f$ содержит 12264 документа.

Первый уровень иерархии

Φ^1 инициализировалась, как описано в алгоритме. Применялись регуляризаторы декорреляции терминов нефоновых тем с коэффициентом, увеличенным пропорционально увеличению размера коллекции и сглаживания Φ для фоновой темы.

На рис.4 изображена визуализация полученной модели. Видно, что полученная модель отличается от первого уровня иерархии ПостНауки незначительно. Таким образом, добавленные документы из коллекции Хабрахабра распределились по существующим темам.

Рис. 4.3: Визуализация первого уровня иерархии

Второй уровень иерархии

Φ^2 инициализировалась, как описано в алгоритме.

На втором уровне иерархии использовалась та же стратегия регуляризации, что и для второго уровня модели базовой коллекции. Коэффициент при декоррелирующем регуляризаторе был также изменен пропорционально размеру коллекции.

В таблице 4.2 приведены темы построенной модели с соответствующими им количеством и долей статей Хабрахабра. Таблица содержит все темы с наибольшим количеством статей Хабрахабра и несколько примеров тем с малой долей статей Хабрахабра. Как и ожидалось, добавленные документы распределились по темам, характерным для

Название темы t	$\lvert D_t \rvert$	P_t
'технологии', 'интернет', 'социальные сети'	6006	0.95
'математика', 'статистика', 'нейронные сети'	2434	0.93
'экономика', 'сша', 'япония'	2203	0.90
'право', 'юриспруденция', 'закон'	1463	0.92
'образование', 'наука', 'университет'	1298	0.86
'философия', 'история философии', 'политическая философия'	142	0.34
'медицина', 'биомедицина', 'онкология'	120	0.28
'биология', 'эволюция', 'днк'	111	0.15
'история', 'история россии', 'средневековье'	94	0.15

Таблица 4.2: Распределение новых документов по темам. $\lvert D_t \rvert$ – количество документов Хабрахабра, отнесенных к теме t по порогу вероятности $p(t|d) > 0.1$. P_t – доля документов Хабрахабра среди всех документов темы t

Хабрахабра. При этом темы, характерные только для ПостНауки, содержат значительно меньше статей Хабрахабра. Так как объединенная коллекция содержит в основном документы Хабрахабра, такое распределение приводит к существенным различиям в размере тем, тогда как при базовом подходе к построению модели все темы имеют примерно одинаковый размер.

4.5 Сравнение алгоритмов

Предлагаемый метод построения тематической иерархии имеет несколько существенных отличий от базового метода: выбор множества добавляемых документов с помощью фильтрации, использование словаря базовой коллекции, инициализация начальных значений параметров модели. Для исследования того, как эти отличия влияют на качество иерархии, приведем сравнительный анализ нескольких модификаций предлагаемого алгоритма и базового алгоритма. Будем оценивать качество каждого алгоритма по метрикам качества иерархий, предложенным в секции 3.4.

Будем сравнивать модификации алгоритмов на одном и том же размере объединенной коллекции D, то есть в каждом случае в базовую коллекцию добавляется одинаковое количество документов D_1. Каждая рассматриваемая модификация алгоритма может использовать или не использовать фильтрацию (**Ф**), сокращение словаря до словаря базовой коллекции (**С**) и инициализацию параметров (**И**). Тогда опишем

в таком виде исследуемые модификации.

- **Ф- С- И-**. Это базовый алгоритм. В качестве D_f выбирается случайное подмножество статей D_1 и строится модель объединенной коллекции $D = D_0 \cup D_f$.
- **Ф+ С- И-**. Для получения D_f используется фильтрация, описанная в секции 4.3.1, далее строится модель объединенной коллекции $D = D_0 \cup D_f$.
- **Ф- С+ И-**. Предварительно находится словарь базовой коллекции. В качестве D_f выбирается случайное подмножество статей D_1. Модель строится только по словам из словаря D_0.
- **Ф- С- И+**. Предварительно строится модель базовой коллекции. В качестве D_f выбирается случайное подмножество статей D_1. Модель объединенной коллекции $D = D_0 \cup D_f$ строится с инициализацией параметрами начального приближения.
- **Ф- С+ И-**. Предварительно находится словарь базовой коллекции. Для получения D_f используется фильтрация, описанная в секции 4.3.1. Модель строится только по словам из словаря D_0.
- **Ф+ С+ И+-**. Это итеративная модификация предлагаемого алгоритма, в которой новые документы добавляются в коллекцию малыми порциями итеративно . При этом на каждой итерации за начальное приближение берется модель с предыдущей итерации. Таким образом, на каждой итерации матрица для инициализации параметров отличается от предыдущей.
- **Ф- С+ И+**. В качестве D_f выбирается случайное подмножество статей D_1. Для инициализации используется модель базовой коллекции, модель строится только по словам словаря D_0.
- **Ф+ С+ И+**. Это предлагаемый алгоритм.

На графике 4.4 приведено сравнение среднего значения метрики S_{emb} для каждого из описанных алгоритмов, как предложено в секции 3.4.1. На графике 4.5 эти же алгоритмы сравниваются с помощью метрик качества ранжирования, описанных в секции 3.4.2, с использованием метрики S_{emb}. Во всех алгоритмах гиперпараметры моделей, стратегия регуляризации и размер добавляемой коллекции были такими же, как в вычислительном эксперименте, описанном в секции 4.4. В итеративном алгоритме на каждой итерации в коллекцию добавлялось

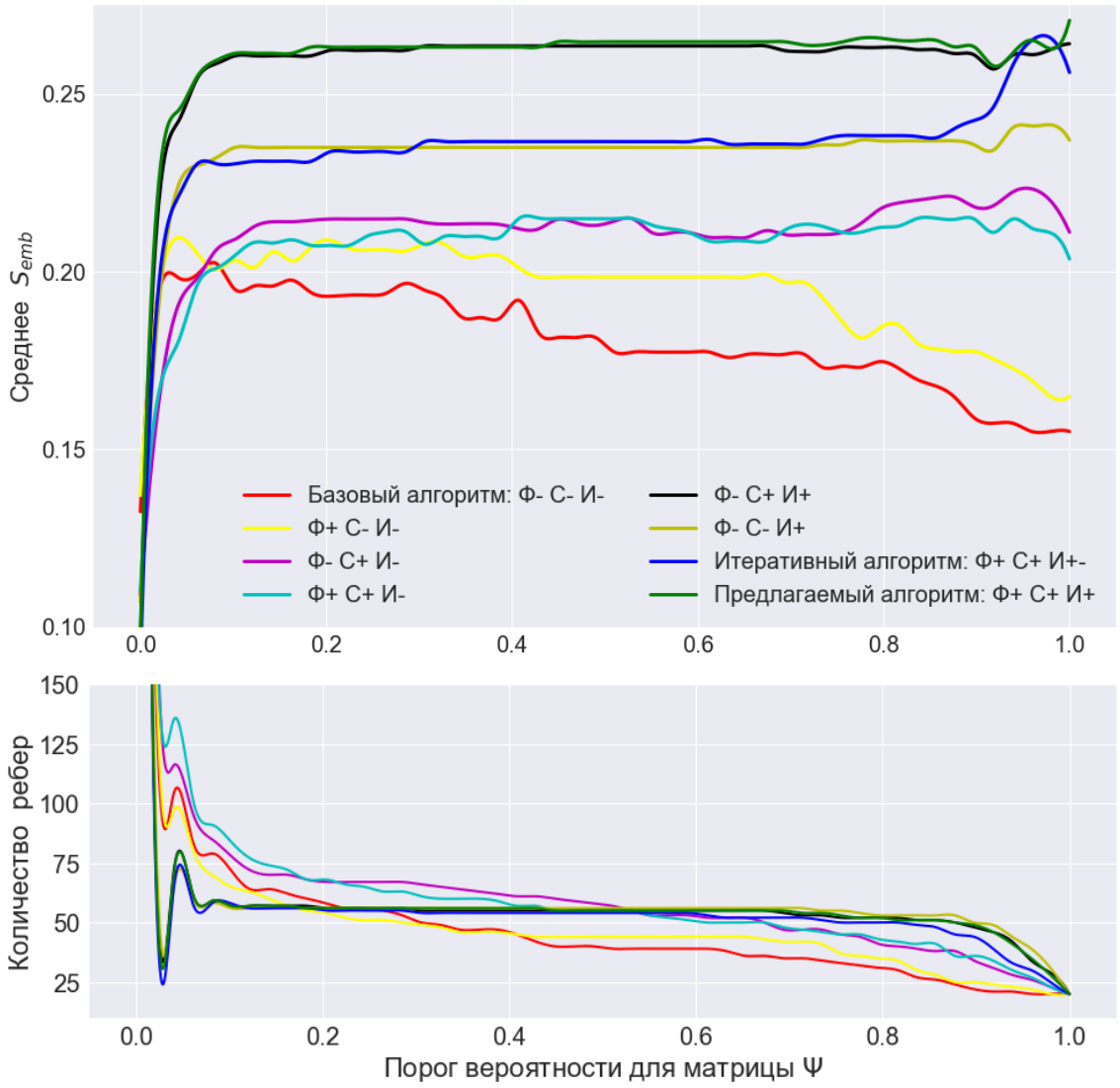

Рис. 4.4: Сравнение среднего качества ребер иерархии по метрике S_{emb} для исследуемых алгоритмов.

количество документов равное 10% от размера коллекции на предыдущей итерации. Таким образом для добавления всей выборки D_f понадобилось 19 итераций.

На графике 4.4 стоит рассматривать области низких и высоких порогов вероятности отдельно. В области высоких порогов усреднение метрики идет только по нескольким самым вероятным ребрам в иерархии, то есть по тем, которые модель считает наилучшими. В области низких порогов усреднение идет по большому количеству ребер. Для наглядности на графике 4.4 снизу изображены зависимости количества ребер в иерархиях от порога. Видно, что для алгоритмов с лучшим качеством количество ребер изменяется слабо в широкой области порогов вероятности. Это можно считать показателем качества иерархии, так как в таком случае иерархия разрежена и почти все ненуле-

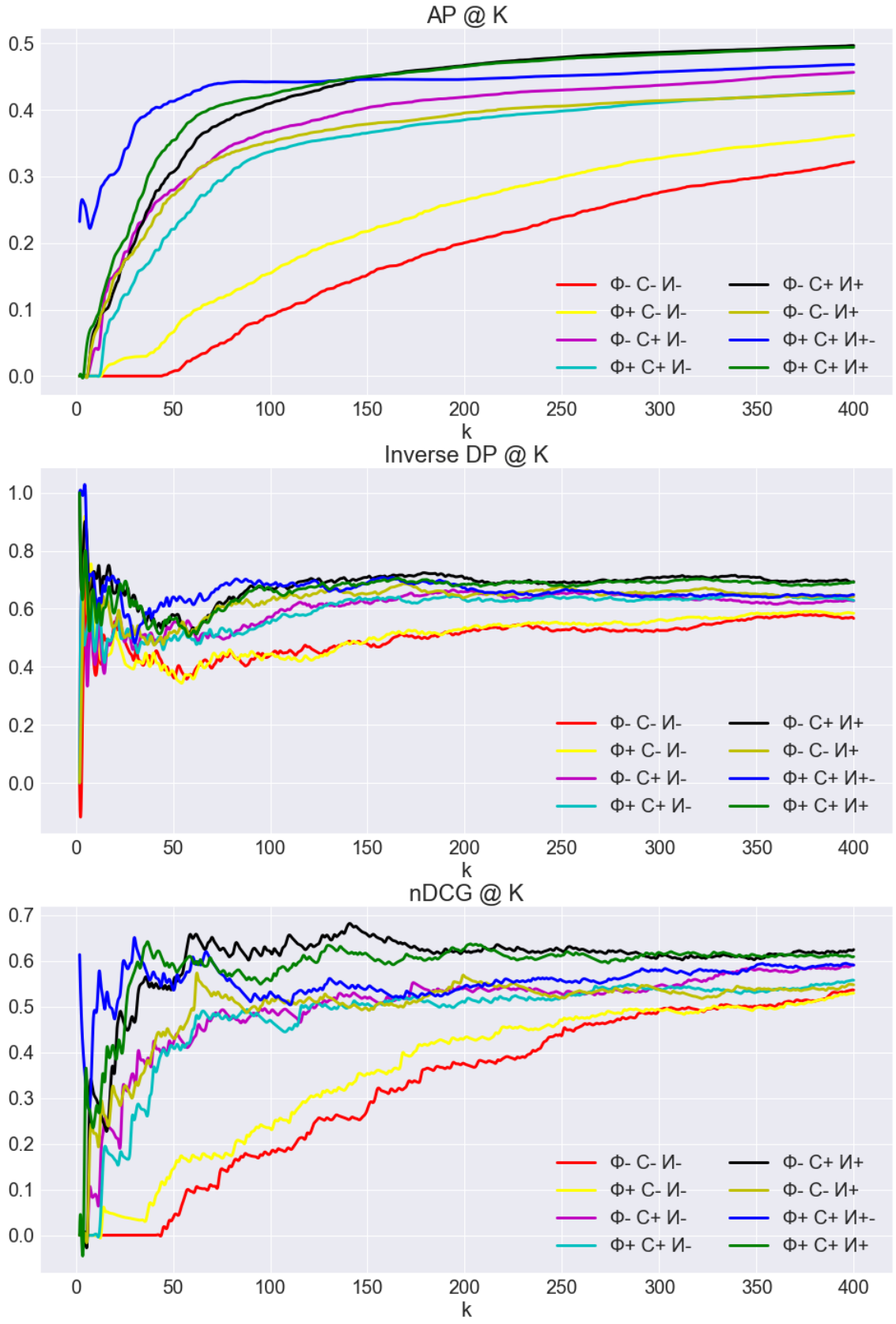

Рис. 4.5: Сравнение алгоритмов по метрикам качества ранжирования. Верное ранжирование задает метрика S_{emb}.

вые значения значительно больше нуля. Если среднее качество падает с увеличением порога, как для базового алгоритма, это говорит о том, что алгоритм плохо ранжирует ребра иерархии. Это подтверждается метриками качестве ранжирования на графике 4.5.

Из приведенных результатов видно, что все три модификации базового алгоритма (фильтрация, словарь и инициализация) по отдельности улучшают качество иерархии. При этом наибольший прирост качества дает хорошая инициализация модели, меньший прирост дает сокращение словаря и меньше всего влияет на качество модели фильтрация. Видно, что фильтрация по tf-idf близости существенно не меняет модель, которая уже использует отфильтрованный словарь. Однако, используемые метрики качества говорят только о качестве представления тем и связей между темами их топ-токенами. В случае модели без фильтрации в выборку попадают документы, которые плохо тематизируются и не относятся ни к одной из нефоновых тем со значимой вероятностью. Хотя это может не влиять значительно на качество модели по используемым метрикам, это нежелательный эффект, с которым помогает справится фильтрация.

Наилучшее качество показывает предлагаемый алгоритм, почти такое же качество модели получается без фильтрации. Если использовать алгоритм итеративно с изменением параметров инициализации, то качество ухудшается, то есть лучшей стратегией является инициализация начальным приближением даже в случае, когда данные приходят порциями и алгоритм необходимо применять много раз.

Глава 5

Заключение

В этой работе было предложено несколько автоматических метрик качества для отношений «родитель-ребенок» в иерархической тематической модели. Было показано, что метрика качества S_{emb}, основанная на векторных представлениях слов, хорошо аппроксимирует мнения асессоров о том, существует ли связь между темами. Другие метрики продемонстрировали меньшую, но приемлемую согласованность с асессорами.

Кроме того, были предложены два подхода для измерения качества иерархии в целом, основанные на усреднении качества ребер и на качестве ранжирования ребер, которое задает модель.

Для решения задачи агрегирования было предложено несколько модификаций базового метода построения моделей гетерогенных текстовых коллекций: фильтрация, сокращение словаря, инициализация. Было проведено сравнение качества моделей, построенных с использованием этих модификаций. Показано, что предлагаемый алгоритм, использующий все три модификации, дает наилучшее качество модели.

Результаты работы были представлены на 24-ой международной конференции по компьютерной лингвистике и интеллектуальным технологиям (”Диалог”2018) и на 60-й Научной конференции МФТИ.

Литература

[1] Laura A. Sill, "Introduction to Cataloging and Classification", ***Library Collections, Acquisitions, and Technical Services***, vol. 31, no. 2, pp. 110–111, jun 2007.

[2] David M. Blei and John D. Lafferty, "A correlated topic model of Science", ***The Annals of Applied Statistics***, vol. 1, no. 1, pp. 17–35, jun 2007.

[3] A. J. B. Chaney and D. M. Blei, "Visualizing Topic Models.", ***Icwsm***, 2012.

[4] Jason Chuang, Christopher D. Manning, and Jeffrey Heer, "Termite : Visualization Techniques for Assessing Textual Topic Models", ***Proceedings of the International Working Conference on Advanced Visual Interfaces - AVI '12***, 2012.

[5] David M. Blei, "Probabilistic topic models", ***Commun. ACM***, vol. 55, no. 4, pp. 77–84, 2012.

[6] 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.

[7] David M. Blei, Andrew Y. Ng, and Michael I. Jordan, "Latent dirichlet allocation", ***Journal of Machine Learning Research***, vol. 3, pp. 993–1022, 2003.

[8] Michal Rosen-Zvi, Thomas L. Griffiths, Mark Steyvers, and Padhraic Smyth, "The author-topic model for authors and documents", ***CoRR***, vol. abs/1207.4169, 2012.

[9] Chaitanya Chemudugunta, Padhraic Smyth, and Mark Steyvers, "Modeling general and specific aspects of documents with a probabilistic topic model", in ***NIPS***, Bernhard Schölkopf, John C. Platt, and Thomas Hofmann, Eds. 2006, pp. 241–248, MIT Press.

[10] Khoat Than and Tu Bao Ho, "Fully sparse topic models", in ***ECML/PKDD (1)***, Peter A. Flach, Tijl De Bie, and Nello Cristianini, Eds. 2012, vol. 7523 of ***Lecture Notes in Computer Science***, pp. 490–505, Springer.

[11] 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.

[12] 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.

[13] K. V. Vorontsov, "Additive regularization for topic models of text collections", ***Doklady Mathematics***, vol. 89, no. 3, pp. 301, May 2014.

[14] Konstantin Vorontsov and Anna Potapenko, "Additive regularization of topic models", ***Machine Learning***, vol. 101, no. 1-3, pp. 303, October 2015.

[15] D. M Blei, T. Griffiths, Michael I. Jordan, and J. Tenenbaum, "Hierarchical topic models and the nested chinese restaurant process", 2003.

[16] 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.

[17] N. A. Chirkova and K. V. Vorontsov, "Additive regularization for hierarchical multimodal topic modeling", 2016.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] Gerlof Bouma, "Normalized (pointwise) mutual information in collocation extraction", ***Proceedings of GSCL***, pp. 31–40, 2009.

[24] 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.

[25] 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.

[26] C. Chemudugunta, P. Smyth, and M. Steyvers, "Modeling general and specific aspects of documents with a probabilistic topic model", ***Nips***, 2006.

[27] 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.

[28] 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.

[29] 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***, vol. 12, pp. 2749–2775, 2011.

[30] Jay Pujara and Peter Skomoroch, "Large-Scale Hierarchical Topic Models", ***NIPS Workshop on Big Learning***, 2012.

[31] 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.

[32] Hinrich Schütze and Jan Pedersen, "A Vector Model for Syntagmatic and Paradigmatic Relatedness", ***Making Sense of Words: Proceedings of the Conference***, 1993.

[33] 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.

[34] Li Hang, "A Short Introduction to Learning to Rank", ***IEICE Transactions on Information and Systems***, 2011.

[35] Anastasia Ianina, Lev Golitsyn, and Konstantin Vorontsov, "Multi-objective topic modeling for exploratory search in tech news", in ***Conference on Artificial Intelligence and Natural Language***. Springer, 2017, pp. 181–193.