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

# Todo list

# Глава 1

# Введение

# 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) [1]. В ней матрицы $\Phi$ и $\Theta$ находятся с помощью максимизации логарифма правдоподобия:

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

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) [2], в которой вектора $\phi_t$ и $\theta_d$ генерируются из распределений Дирихле. Кроме того, было предложено множество обобщений LDA, учитывающих дополнительные требования к модели [3, 4, 5].

### 2.2.3 Аддитивная регуляризация тематических моделей

Априорным распределениям параметров тематической модели соответствуют регуляризаторы в задаче оптимизации. Аддитивная регуляризация тематических моделей (модель ARTM) [6] 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 $\tau_i$:

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

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

Как показано в [6], 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 $\phi_t$ 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 дивиргенции Кульбака-Лейблера [6]:

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

где $\beta_0$ и $\alpha_0$ — заданные положительные коэффициенты. Введение этого регуляризатора, как показано в [6], 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 [6]:

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

### 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 $\Phi^m$ — соответсвующие 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 $R_i$ [6]:

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

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

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

С другой стороны, модель иерархического распределения патинко (hPAM) [8] представляет собой направленный ациклический многодольный граф. Модель иерархического процесса Дирихле (hHDP) [9] 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 [10, 11].

### 2.3.2 Иерархическая модель ARTM

В иерархической модели ARTM (hARTM) [12] 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 [12]:

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

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

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

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 $\Psi$ образует $|A|$ дополнительных столбцов матрицы $\Theta$, 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 [13] 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 [14].

В работе [15] 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 [16]. 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 [14].

Другой тип метрик качества основан на поточечной взаимной информации (PMI). Следуя [17], рассмотрим три метрики этого типа.

В [18] 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 [19] 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 [13] 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 [13] и [16].

В [16] 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 [20]. 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 [14]:

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

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 [13], 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 [16], 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 [13, 16], 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 краудсорсинг 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 |
|---|---|
| $\text{S}_{emb}$ | 0.878 |
| $\text{S}_{Hell}$ | 0.815 |
| $\text{S}_{KL}$ | 0.790 |
| $\text{S}_{coh}$ | 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 $\text{S}_{emb}$ (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 $\text{S}_{emb}$, так как она показала наилучшую согласованность.

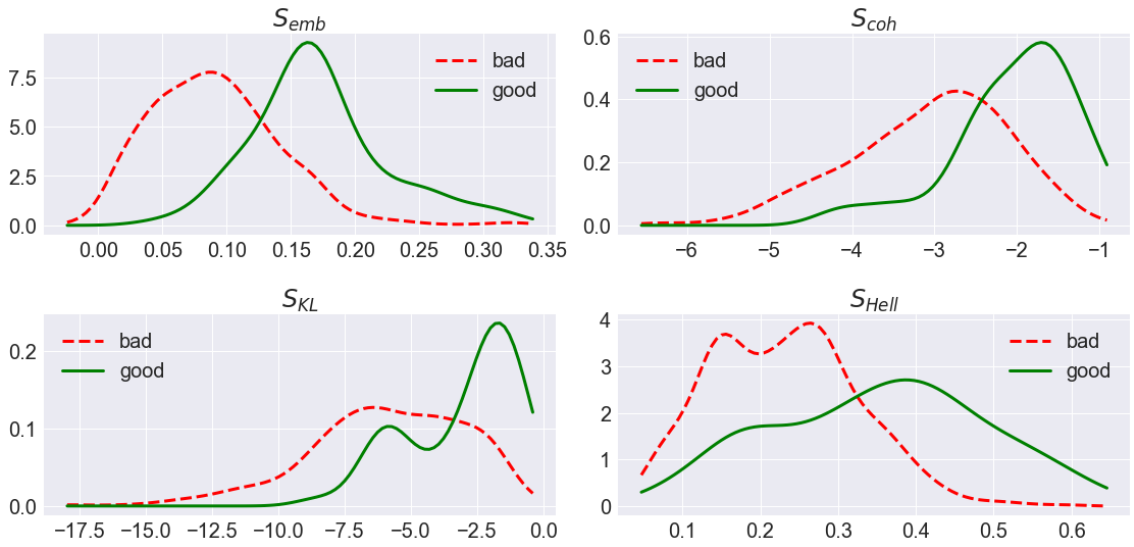


Рис. 3.1: Вероятностные распределения значений метрик для «хороших» и «плохих» ребер.

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».

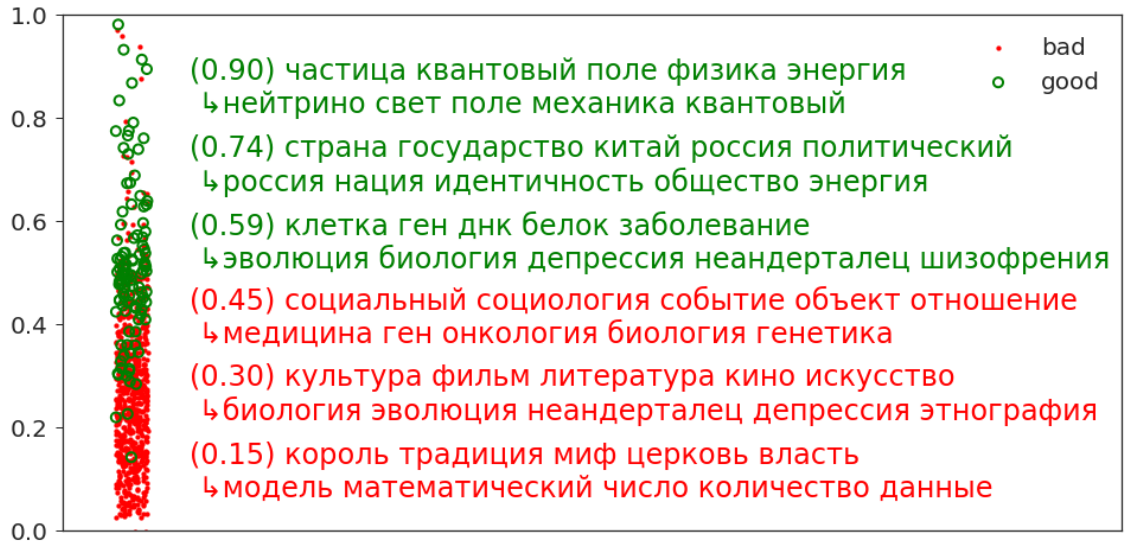


Рис. 3.2: Примеры пар тем из асессорского эксперимента с соответствующими им значениями метрики $S_{emb}$ и вердиктом асессоров. Каждая тема и подтема представлены своими 5 топ-токенами.

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 [13] 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 $\Psi$:

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

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 $\Psi$. 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 $\Psi$ с помощью принятых метрик качества ранжирования, таких как:

- Средняя точность (Average Precision, AP@k) – описана в [21].
- Inverse Defect Pairs, IDP@k – обратное к значению количества пар, которые ранжируются алгоритмом в неправильном порядке.
- Normalized Discounted Cumulative Gain, nDCG@k – описана в [21].

# 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.

| 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 |

Таблица 4.1: Доля статей Хабрахабра в темах модели объединенной коллекции

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 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.

### 4.3.1 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. 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 $\boldsymbol{d_n} = [\text{tf}(\text{term}_i, d_n) \cdot \text{idf}(\text{term}_i, D)]_{i=1}^{|W|}$, 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:

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

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

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

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

### 4.3.2 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 $\Phi^1_{\text{new}}$ первого уровня иерархии, соответсвующие токенам старого словаря, не должны значительно отличаться от соответствующих строк старой матрицы $\Phi^1$. Поэтому инициализируем эту подматрицу $\Phi^1_{\text{new}}$ матрицей $\Phi^1$, а остальные строки инциализируем 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 $\Phi^2_{\text{new}}$, 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 разрежива-ния $\Theta$ по новым темам для документов старой коллекции.

## 4.4 Сравнение алгоритмов

# Глава 5

# Заключение

# Литература


[1] 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.

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

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

[4] 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.

[5] 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.

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

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

[8] 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.

[9] 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.

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

[11] 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.

[12] NA Chirkova and KV Vorontsov, "Additive regularization for hierarchical multimodal topic modeling", 2016.

[13] 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.

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

[15] 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.

[16] 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.

[17] 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.

[18] 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.

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

[20] 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.

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