Московский государственный университет имени М.В. Ломоносова
Факультет вычислительной математики и кибернетики
Кафедра математических методов прогнозирования

Чиркова Надежда Александровна

Иерархические тематические модели для интерактивной навигации по коллекциям текстовых документов

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

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

Москва, 2016

Аннотация

Иерархический тематический навигатор — это система, помогающая пользователю ориентироваться в коллекции текстовых документов и показывающая структуру коллекции в виде иерархии тем. Находясь в любой теме, пользователь выбирает, в какие подтемы или надтемы перейти; так он быстро находит множество интересных ему документов. Подобные системы удобно создавать на основе тематических моделей, представляющих каждый документ коллекции в виде смеси тем и для каждой темы определяющих ее лексических состав. Для этого необходимо усложнить модель, чтобы она могла описывать связи надтема-подтема. Цель данной работы — разработать подход к построению иерархических тематических моделей, которые станут удобной основой для создания тематического навигатора. Введены формальные требования, которым должна удовлетворять модель. Предложено два способа построения иерархии: нисходящий, то есть постепенно дробящий темы на подтемы, и строящий всю тематическую иерархию целиком. Предложены критерии качества, оценивающие выполнение различных свойств иерархии. В экспериментах показывается, что предложенные способы позволяют строить интерпретируемые тематические иерархии, в которых удобно ориентироваться пользователю. На основе одной из предложенных иерархических моделей создан интерактивный навигатор по коллекции статей двух русскоязычных конференций по анализу данных.

Содержание

1 Введение

Людям часто приходится выбирать источники информации. В эпоху информационных технологий у нас появляется доступ к неограниченному объему знаний, доступных через сеть Интернет, но физически мы можем ознакомиться только с малой частью этих данных. Возникает потребность в системах автоматической организации информации для пользователя. Человечество разработало качественные решения этой проблемы для случаев, когда пользователя знает, какую конкретно информацию хочет получить. Однако вопрос, как найти полезную информацию, не задавая поискового запроса, до сих пор остается открытым.

Решением могут служить иерархические тематические навигаторы. По аналогии с идеей файловой системы компьютера навигатор позволяет организовать документы большой информационной коллекции в виде иерархии тем. Спускаясь по иерархии сверху вниз, пользователь выбирает интересные ему подтемы и видит документы, релевантные текущей теме. Таким образом, пользователь быстро находит узкую группу материалов, которые ему интересны, изначально не зная, какова тематика всей коллекции. Помимо организации документов, такая система должна без участия человека аннотировать темы, то есть кратко характеризовать их содержание, чтобы пользователь мог быстро отличать темы друг от друга.

Последние пятнадцать лет активно развивается раздел машинного обучения, решающий задачу поиска тем в коллекции документов — вероятностное тематическое моделирование. Тематические модели описывают каждый документ как смесь тем, а темы представляют в виде дискретного вероятностного распределения над множеством терминов. Иными словами, модель задает компактное представление для коллекции, которое позволяет быстрее ознакомиться с ее содержанием. Однако на больших текстовых коллекциях, когда число тем становится равным нескольким сотням или тысячам, даже такое представление в виде набора тем перестает быть удобным. Появляется потребность в построении иерархических тематических моделей, в которых крупные темы постепенно дробятся на более узкие, специализированные темы.

Таким образом, иерархические тематические модели, во-первых, помогают представить структуру коллекции текстовых документов в виде иерархии тем, а во-вторых, характеризуют лексический состав каждой темы. Это позволяет пользователю наиболее полно познакомиться с областью знаний, к которой относится коллекция. В некоторых исследованиях задача выделения тематической лексики является доминирующей, тогда специально подбирают набор данных, релевантный рассматриваемой области, и строят его тематическую модель.

В литературе широко освещена тема автоматической организации документов в единую иерархию, и разные работы отличаются, во-первых, требованиями, которые авторы предъявляют к иерархии, а во-вторых, математическим аппаратом, который используется для решения задачи.

Большинство подходов к построению иерархий вероятностные: в них термины, темы и документы считаются случайными величинами, а коллекция моделируется с помощью процесса порождения слова в документе. Одна и первых таких иерархических моделей предложена в [11]: иерархия представляется в виде дерева тем, и ее можно достраивать при добавлении новых документов в коллекцию. В [15] ключевая идея состоит в отказе от ограничения на граф: иерархия является многодольным графом, то есть темы могут иметь несколько надтем. Авторы [28] также представляют иерархию в виде многодольного графа и описывают две модели, которые автоматически определяют количество тем и количество уровней, или долей в графе. Одна модель строит иерархию документов, другая — иерархию терминов, совместить эти модели в одной не предлагается. Аналогично, в [19] темы описываются только лексикой, то есть связь документов и тем не моделируется; ключевая особенность — темы представляются как список фраз, а не отдельных терминов, в результате повышается интерпретируемость тем. Список терминов родительской темы получается объединением списков терминов дочерних тем. Этот подход развивается в [9], где модель учитывает не только текстовую, но и иную информацию, представленную в коллекции: авторов, метки времени, локации на карте и т. д. В [22] делают акцент на трех приоритетах: масштабируемость, то есть быстрое построение модели на больших коллекциях, устойчивость, то есть построение похожих моделей при повторных запусках, и интерпретируемость. В [24] к этому списку добавляется еще одна цель: возможность учитывать указания эксперта, например указание объединить две темы. На важность масштабируемости алгоритма обучения также указывают авторы [20].

С другой стороны, разработано много невероятностных подходов к структуризации текстовой информации. К примеру, в [6] авторы ставят задачу систематизации лексики, специфичной для конкретной области знаний, и решают ее с помощью иерархической кластеризации множества ключевых слов этой области. Другой подход, основанный на иерархической кластеризации, но уже документов, описан в [13]; предложен алгоритм, позволяющий добавлять в иерархию новые документы без повторной обработки старых. Авторы [10] предлагают для составления иерархии документов сопоставить каждому документу несколько статей Википедии, а затем выделить ту часть существующей иерархической категоризации открытой энциклопедии, которая соответствует тематике коллекции. Основной акцент в этой статье делается на удобстве пользователя: иерархия должна помогать пользователю выбирать документы для просмотра, давать обзор всей коллекции,

показывать краткую информацию о документах.

Мы видим, что в каждой статье по-своему выбирают свойства, которым должна удовлетворять иерархия, и исходя из них определяют ее структуру. Иерархия может быть деревом или многодольным графом с увеличивающимся числом тем на каждом уровне; документы могут располагаться только в вершинах нижнего уровня или в вершинах промежуточных уровней тоже, и аналогично с терминами; модель может учитывать только лексику предметной области или только релевантные ей документы, а может более полно описывать область, указывая не только термины и документы, характеризующие тему, но и другие объекты, например авторов или географические локации. Полнота описания обычно характерна для вероятностных подходов, потому что в них проще учесть объекты разной природы. Однако в них часто делают акцент на технических, математических или лингвистических особенностях модели. Если ставить задачу построения тематического навигатора, который будет удобен пользователю, то нужно в первую очередь анализировать, какие свойства иерархии упростят его работу с коллекцией. Именно так ставят задачу в [10].

Цель данной работы — предложить способ построения тематической иерархии, которая наиболее полно описывает коллекцию и учитывает всю информацию, известную о документах, и на основе которой будет удобно создавать тематический навигатор для пользователя. Отметим, что понятия тематической иерархии и тематического навигатора не тождественны: иерархия — это математическая модель коллекции документов, а навигатор — это способ представления информации для пользователя. В данной работе предполагается, что навигатор составляется на основе модели путем извлечения из нее наиболее важной информации, полученной из коллекции в результате обучения.

Перечислим и обоснуем требования, которым должны удовлетворять иерархическая тематическая модель и тематический навигатор:

- требование полноты описания: иерархия должна быть представлена в виде графа тем, и каждая тема должна описываться, как минимум, релевантными терминами и документами. Если для документа известно иное признаковое описание, кроме текстового, то можно выделять также списки признаков, релевантных теме. Это требование позволяет учитывать как можно больше информации, содержащейся в коллекции. Отметим, что документы могут описываться несколькими темами, располагающимися на одном или разном уровнях, то есть документ не обязан быть монотематическим.

- структурное требование: иерархия должна быть представлена в виде разреженного многодольного графа с увеличивающимся количеством тем на каждом уровне. Иными словами, темы могут иметь несколько надтем, но в целом граф должен быть

близок к дереву. В таком графе пользователю будет проще переходить в смежные темы, и он лучше познакомится с коллекцией.

- требование разреженности: каждый термин и документ должен быть релевантен небольшому количеству тем в рамках одного уровня. Пользователю сложно воспринимать длинный список объектов, характеризующих тему.

- требование масштабируемости: обучение модели должно требовать малого числа проходов по коллекции. В противном случае область применимости метода ограничится небольшими коллекциями.

Кроме того, можно выделить несколько желаемых, но необязательных требований к модели:

- требование расслоения словаря: термин, характеризующий тему, должен иметь практически нулевые вероятности появления в ее подтемах. Так будет проще видеть различия между подтемами. Можно сформулировать ослабленное требование к навигатору, получающемуся из модели: термин, характеризующий тему, не должен показываться в ее подтемах.

- требование расслоения документа: для каждого документа хочется знать, насколько узка его тематика, то есть видеть, какая доля документа сконцентрирована на каждом уровне иерархии. Модель должна позволять оценивать этот вектор пропорций для каждого документа в процессе обучения.

Мы будем рассматривать только вероятностные подходы, потому что они являются обобщениями классических тематических моделей и позволяют выполнить требование полноты описания.

В следующем за введением обзоре литературы мы проанализируем, каким требованиям не удовлетворяют существующие подходы. Далее в работе вводятся и экспериментально исследуются два предложенных подхода к построению иерархии, а также описывается способ построения тематического навигатора на основе этих моделей. Затем производится сравнение двух предложенных подходов.

2 Обзор литературы

2.1 Обозначения и определения

Пусть D – множество текстовых документов, W – множество всех употребляемых в коллекции терминов — слов или словосочетаний, также именуемое словарем. Каждый до-

кумент $d \in D$ представляет собой последовательность n_d терминов $(w_1, \ldots, w_{n_d})$, принадлежащих словарю W. В тематическом моделировании принимается гипотеза о том, что порядок терминов в тексте не важен для определения его тематики. Тогда коллекцию можно представить в виде матрицы частот слов с элементами n_{dw} – частота вхождения термина w в документ d. Длину документа будем обозначать n_d. По матрице частот слов можно оценить вероятности появления терминов в документе: $p(w|d) = \frac{n_{dw}}{n_d}$.

Если документ, помимо текстового описания, характеризуется другими признаками, например автором или ключевыми словами, то говорят, что в него входят элементы разных модальностей. Каждой модальности соответствует отдельный словарь, состоящих из всевозможных значений элементов модальности. В примере выше словари авторов и ключевых слов будут состоять соответственно из фамилий всех авторов, написавших документы, и всех ключевых слов, встретившихся в коллекции. В этом случае считается, что термины также формируют отдельную, текстовую, модальность. Множество модальностей будем обозначать M, а их словари — W^m, $m \in M$; $W = \cup_{m \in M} W^m$.

Как уже отмечено во введении, тематической иерархией мы называем многодольный граф с вершинами-темами и ребрами, показывающими связи между ними. Доли графа будем считать упорядоченными в порядке роста количества тем-вершин в каждой доле и называть их уровнями или слоями. Если две темы соединены ребром, то тему, находящуюся на уровне выше, называют родительской, а тему на следующем уровне — дочерней темой. Если самый верхний уровень иерархии представлен одной темой, ее называют корневой. Уровни иерархии будем обозначать $1, \ldots, L$, множества тем этих уровней — $S_1, \ldots, S_l$. Обычные, не иерархические, тематические модели, в которых все темы равносильны и формируют одно множество тем S, называют плоскими.

Будем говорить, что в отличие от иерархии-дерева, иерархия в виде многодольного графа разрешает множественное наследование тем, то есть допускает ситуацию, когда одна тема имеет несколько родительских.

2.2 Плоские тематические модели и их обобщения

Вероятностный латентный семантический анализ. Первая вероятностная тематическая модель была предложена Томасом Хоффманом в [12]. В ней вводятся две матрицы параметров: $\Phi = \{\phi_{ws}\}_{w \in W, s \in S}$, $\phi_{ws} = p(w|s)$ и $\Theta = \{\theta_{sd}\}_{s \in S, d \in D}$, $\theta_{sd} = p(s|d)$. Будем обозначать дискретные распределения $\phi_s = \{\phi_{ws}\}_{w \in W}$ и $\theta_d = \{\theta_{sd}\}_{s \in S}$. Модель предполагает следующий процесс генерации текстовой коллекции документов:

1: **для всех** документов $d \in D$:
2: **для** $i = 1, \ldots, n_d$ сгенерировать i-е слово w в документе:

3: сгенерировать $s \sim \text{Mult}(\theta_d)$;

4: сгенерировать $w \sim \text{Mult}(\phi_s)$.

Здесь и далее $\text{Mult}(x)$ обозначает дискретное распределение, задаваемое вектором вероятностей x.

Этому процессу соответствует следующая модель коллекции:

$$p(w|d) = \sum_{s \in S} p(w|s)p(s|d) = \sum_{s \in S} \phi_{ws}\theta_{sd}.$$

При этом вводится *гипотеза условной независимости*:

$$p(w|s,d) = p(w|s),$$

вероятность появления элемента w в теме s не зависит от того, к какому документу он относится.

Таким образом, коллекция, представленная матрицей частот слов $F = \{n_{dw}\}_{d \in D, w \in W}$, моделируется матричным произведением:

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

В этом подходе предполагается, что каждое слово в документе связано с некоторой скрытой, или латентной, темой; тогда величины $p(s|d,w)$ — это скрытые переменные.

Параметры модели настраиваются методом максимального правдоподобия с помощью EM-алгоритма:

$$\sum_{d \in D} \sum_{w \in W} n_{dw} \ln \sum_{s \in S} \phi_{ws}\theta_{sd} \to \max_{\Phi,\Theta}, \tag{1}$$

$$\sum_{w \in W} \phi_{ws} = 1; \phi_{ws} \geqslant 0; \quad \sum_{s} \theta_{sd} = 1; \theta_{sd} \geqslant 0. \tag{2}$$

Этот подход получил название вероятностный латентный семантический анализ (Probabilistic Latent Semantic Analysis, PLSA).

Латентное размещение Дирихле В 2003 году Дэвид Блей в [7] предложил латентное размещение Дирихле (Latent Dirichlet Allocation, LDA) — байесовскую трактовку модели PLSA. В ней вводятся априорные распределения Дирихле на параметры модели. Обозначим гиперпараметры априорных распределений α, η, ξ. Модель основана на следующем процессе генерации документа d:

1: сгенерировать длину документа из пуассоновского распределения: $n_d \sim \text{Poisson}(\xi)$

2: сгенерировать распределение над множеством тем $\theta_d \sim \text{Dir}(\alpha)$

3: **для** $i = 1, \ldots, n_d$ сгенерировать i-е слово w в документе:

4: сгенерировать $s \sim \text{Mult}(\theta_d)$;

5: сгенерировать $w \sim \text{Mult}(\phi_s)$.

Позже в модель добавили еще один шаг генерации вероятностных распределений тем, который выполняется до генерации документов:

сгенерировать распределение $\phi_s \sim \text{Dir}(\eta)$ для всех $s \in S$.

Обучение модели состоит в поиске апостериорных распределений на параметры Φ и Θ. Для этого используют вариационный вывод [7] или семплирование Гиббса [26].

Обобщения LDA Модель LDA стала основой для многочисленных обобщений, позволяющих учитывать в тематической модели дополнительную информацию или предъявлять к ней дополнительные требования. Рассмотрим некоторые обобщения, связанные с введенными требованиями к модели.

В модели Author Topic Model (ATM) [5] предлагается способ учета авторов документов в модели. Теперь каждый документ ассоциируется с равномерным распределением над множеством его авторов $a \in A_d$, а каждый автор — с распределением над множеством тем θ_a. Добавляется шаг генерации распределений $\theta_a \sim \text{Dir}(\alpha)$, $a \in A = \cup_d A_d$. Для генерации слова w в документе d сначала выбирается автор a из равномерного распределения $\text{Unif}(A_d)$, затем тема $s \sim \text{Mult}(\theta_a)$, затем слово $w \sim \text{Mult}(\phi_s)$. Модель обучается с помощью семплирования Гиббса.

Несколько подходов разработано для построения разреженных тематических моделей, то есть таких, в которых каждый термин и документ связаны с малым числом тем. В подходе Fully Sparse Topic Models (FSTM) [23] предлагается вместо априорных распределений Дирихле на ϕ_s и θ_d использовать другие априорные распределения, которые повышают разреженность.

В модели Topic Model with Special Words (SWB) [8] вводится предположение, что документ состоит из терминов трех типов: тематическая лексика, специальная лексика документа и общая лексика коллекции. Соответственно, для генерации слова сначала выбирается множество, из которого оно генерируется, а затем слово генерируется либо согласно стандартной процедуре LDA, если слово тематическое, или из одного из двух дискретных распределений, соответствующих общему распределению коллекции или специфическом распределению документа. Апостериорное распределение находится с помощью семлпирования Гиббса.

Недостаток всех этих моделей в том, что их трудно совместить в одной. Поэтому в литературе не встречаются подходы, когда данные обобщения были бы применены к иерархической тематической модели.

Аддитивная регуляризация тематических моделей В [25] предлагается способ совмещения разных модификаций модели в одной благодаря введению регуляризаторов в модель PLSA и поддержке многомодальности. Подход назван аддитивная регуляризация тематических моделей (Additive Regularization of Topic Models, ARTM).

В ARTM вводится несколько матриц тем, по одной на каждую модальность: $\Phi^m = \{p(w|s)\}_{w\in W^m, s\in S}$. Обратим внимание, что столбцы каждой матрицы Φ^m являются дискретными вероятностными распределениями. Обозначим $\Phi = \cup_m \Phi^m$.

Кроме того, вводятся регуляризаторы — дополнительные критерии $R_i(\Phi, \Theta)$, характеризующие качество модели. Например, регуляризаторы могут задавать лингвистические свойства, которыми должна обладать модель.

Для оценивания параметров модели максимизируется логарифм правдоподобия со взвешенной суммой регуляризаторов $R(\Phi, \Theta) = \sum_i \tau_i R_i(\Phi, \Theta)$:

$$\sum_{m\in M} \eta_m \sum_{d\in D} \sum_{w\in W^m} n_{dw} \ln \sum_{s\in S} \phi_{ws}\theta_{sd} + R(\Phi, \Theta) \to \max_{\Phi,\Theta}, \tag{3}$$

$$\sum_{w\in W^m} \phi_{ws} = 1; \phi_{ws} \geqslant 0; \quad \sum_s \theta_{sd} = 1; \theta_{sd} \geqslant 0, \tag{4}$$

где коэффициенты η_m введены для балансирования важности модальностей W^m, а коэффициенты τ_i — для балансирования между оптимизацией нескольких критериев.

В [25] доказана следующая теорема.

Теорема 1 (Воронцов, Потапенко) *Если $R(\Phi, \Theta)$ непрерывно дифференцируема, то стационарная точка задачи (3)–(4) удовлетворяет следующей системе уравнений:*

E-шаг: $p(s|d, w) = \underset{s\in S}{\text{norm}}(\phi_{ws}\theta_{sd})$;

M-шаг: $n_{ws} = \sum_{d\in D} \eta_{m(w)} n_{dw} p(s|d, w), \quad n_{sd} = \sum_{w\in W} \eta_{m(w)} n_{dw} p(s|d, w)$;

$$\phi_{ws} = \underset{w\in W^m}{\text{norm}}\left(n_{ws} + \phi_{ws}\frac{\partial R}{\partial \phi_{ws}}\right), \; \theta_{sd} = \underset{s\in S}{\text{norm}}\left(n_{sd} + \theta_{sd}\frac{\partial R}{\partial \theta_{sd}}\right);$$

где $m(w)$ обозначает модальность сущности w, а оператор norm преобразует вещественный вектор в дискретное распределение: $\underset{t\in T}{\text{norm}}(x_t) = \frac{\max\{x_t,0\}}{\sum_{t\in T}\max\{x_t,0\})}$.

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

В результате можно предъявлять различные требования к модели и учитывать их в виде новых регуляризаторов или модальностей. Недостаток подхода состоит в необходимости настройки коэффициентов τ_i и η_m.

К примеру, требование разреженности можно формализовать в виде регуляризатора разреживания матрицы Φ [25]:

$$R_1(\Phi, \Theta) = R_1(\Phi) = -\tau_1 \sum_{s \in S} \mathrm{KL}(\beta || \phi_s) = C - \tau_1 \sum_{s \in S} \sum_{w \in W} \beta_w \ln \phi_{ws},$$

$\mathrm{KL}(p||q)$ обозначает дивергенцию Кульбака-Лейблера между двумя вероятностными распределениями с одним носителем $x \in X$: $KL(p||q) = \sum_{x \in X} p(x) \frac{\ln p(x)}{\ln q(x)}$, C — константа, не зависящая от Φ, β — заданное распределение над множеством слов, например равномерное или $p(w)$, посчитанное по всей коллекции. В этом регуляризаторе мы стремимся сделать распределения ϕ_s как можно менее похожими на равномерное, то есть поощряем разреженность.

Ввести в модель авторов можно через аппарат модальностей: достаточно составить словарь W^2 всех авторов, которые могут встретиться в коллекции, и добавить матрицу Φ^2 этой модальности к параметрам модели.

Итак, вероятностное тематическое моделирование — это богатый набор методов для семантического анализа текстовых коллекций, который позволяет учитывать разные требования к модели. Благодаря широким возможностям, предоставляемым аппаратом вероятностных моделей, можно усложнять генеративный процесс и с его помощью моделировать иерархические взаимосвязи между темами. Этому посвящен следующий подраздел.

2.3 Иерархические тематические модели

Для построения иерархической тематической модели, как правило, используются два способа: усложнение генеративного процесса порождения коллекции и составление иерархии путем многократного построения плоской тематической модели. При этом возможны три варианта процесса обучения модели: обучение всех уровней одновременно, построение иерархии сверху вниз, когда сначала обучаются параметры крупных тем, а затем специализированных, и в противоположном направлении — построение модели снизу вверх. Последние два способа называют соответственно нисходящим и восходящим способом построения тематической иерархии.

Hierarchical LDA. Первое обобщение LDA для построения иерархий было предложено автором LDA, Дэвидом Блеем, в [11]. Иерархия здесь представляется в виде дерева тем, глубина дерева фиксирована и равна L. Модель сама определяет количество подтем в каждом узле; для этого априорное распределение на путь документа от корня к листу задается с помощью вложенного процесса китайского ресторана (nested Chinese Restraunt Process, nCRP), принимающего в качестве параметров L и коэффициент γ. Пусть мы генерируем путь документа в дереве, в данный момент находимся в вершине s_l уровня l

и хотим выбрать подтему s_{l+1} среди подтем C_l. Тогда мы выберем одну из существующих вершин-подтем с вероятностью, пропорциональной количеству документов, уже относящихся к этой вершине, или с вероятностью, пропорциональной γ, создадим новую подтему. В результате процесс генерации документа коллекции выглядит следующим образом:

1: Начать с корня дерева — вершины s_1.
2: **для всех** $l = 2, \ldots, L$
3: выбрать подтему s_l с помощью nCRP
4: сгенерировать распределение над множеством выбранных подтем $(s_1, \ldots, s_L)$ $\theta_d \sim \mathrm{Dir}(\alpha)$
5: **для** $i = 1, \ldots, n_d$ сгенерировать i-е слово w в документе:
6: сгенерировать уровень $l \sim \mathrm{Mult}(\theta_d)$;
7: сгенерировать $w \sim \mathrm{Mult}(\phi_{s_l})$.

Апостериорное распределение на параметры находится с помощью семплирования Гиббса. Параметры всех уровней иерархии оцениваются одновременно.

Заметим, что каждый новый документ, поступающий в коллекцию, может добавить в дерево новые темы, если его терминология сильно отличается от терминологии существующих тем.

Pachinko allocation, Hierarchical pachinko allocation. В [14] иерархия задается многодольным графом. Каждая вершина s_l ассоциируется с распределением Дирихле $\mathrm{Dir}(\alpha_{s_l})$ над векторами, задающими распределение над темами $s_{l+1} \in S_{l+1}$ следующего уровня. Распределения над множеством слов задаются только в темах последнего уровня. Модель генерации документа d выглядит следующим образом:

1: **для** каждой вершины s_l графа:
2: сгенерировать распределение над темами следующего уровня $\theta_{s_l} \sim \mathrm{Dir}(\alpha_{s_l})$
3: **для** $i = 1, \ldots, n_d$ сгенерировать i-е слово w в документе:
4: сгенерировать путь $(s_1, \ldots, s_L)$ в графе: $s_{l+1} \sim \mathrm{Mult}(\theta_{s_l})$, s_1 — корневая тема
5: сгенерировать $w \sim \mathrm{Mult}(\phi_{s_L})$.

В [15] эта модель усложняется тем, что термины могут генерироваться и из тем промежуточных уровней. К векторам θ_{s_l} добавляется еще один элемент. Если выбирается любая из компонент этого вектора, кроме последней, то слово генерируется из соответствующей темы, иначе выполняется переход на следующий уровень. Первая модель называется Pachinko Allocation, или PAM, а вторая — hPAM.

Модель обучается целиком, без разделения на уровни, с помощью семплирования Гиббса. Основное ее преимущество состоит в учете множественного наследования тем.

Topic Hierarchies of Hierarchical Dirichlet Processes, hHDP. В [28] предлагается модель, похожая на модель PAM. В графе также разрешается множественное наследование, а термины могут генерироваться только из тем последнего уровня. Кроме того, количество тем и количество уровней определяется с помощью иерархического процесса Дирихле (Hierarchical Dirichlet Process, HDP) [27].

Иерархия строится снизу вверх: сначала оценивается количество тем и строится тематическая модель всей коллекции. Затем матрица терминов в темах Φ, обученная на первом этапе, подается на вход HDP в качестве матрицы частот слов, для нее также оценивается количество тем и строится новая тематическая модель, то есть темы нижнего уровня рассматриваются как псевдодокументы для построения модели верхнего уровня. Процесс повторяется, пока количество тем не станет равным 1. Аналогичный процесс предлагается проводить с матрицей документов в темах, которую можно получить из Θ.

Получается, что в отличие от предыдущих двух рассмотренных подходов, мы можем описывать темы либо только терминами, либо только документами.

Scalable Tensor Recursive Orthogonal Decomposition. В [22] предлагается модель, имеющая общие черты в моделями hLDA и PAM, причем модель выбиралась так, чтобы удовлетворить требования масштабируемости, устойчивости и интерпретируемости. Граф вновь представляется деревом тем, и каждая вершина s_l связана с распределением Дирихле над векторами, задающими распределение θ_{s_l} над подтемами этой темы. Процесс генерации документа:

для каждой вершины s_l графа:
    сгенерировать распределение над подтемами следующего уровня $\theta_{s_l} \sim \mathrm{Dir}(\alpha_{s_l})$
для $i = 1, \ldots, n_d$ сгенерировать i-е слово w в документе:
    начать с корневой вершины s_1
    для $l = 1, \ldots, L-1$:
        выбрать подтему $s_{l+1} \sim \mathrm{Mult}(\theta_{s_l})$
    сгенерировать $w \sim \mathrm{Mult}(\phi_{s_L})$.

Для обучения модели используется метод моментов: эмпирические частоты совместной встречаемости терминов приравниваются к теоретическим вероятностям совместного появления терминов в коллекции, и параметры модели оцениваются с помощью тензорных разложений. Обучение выполняется сверху вниз в манере «разделяй и властвуй»: оцениваются параметры темы, затем эта процедура рекурсивно повторяется для ее подтем.

splitLDA. Наконец, наиболее простой рекурсивный подход к построению древесной иерархии был предложен в [20]. В корневой вершине строится модель LDA с множеством тем S_1,

и затем с помощью ее параметров производится разделение коллекции на $|S_1|$ коллекций меньшего размера. Процесс рекурсивно запускается на новых коллекциях.

Преимущество подхода состоит в его высокой масштабируемости, а недостаток — в том, что гиперпараметры LDA, задающие априорные распределения, и количество тем приходится задавать явно в каждой вершине.

2.4 Сравнение рассмотренных подходов

В таблице 1 производится сравнительный анализ подходов с указанием, какие из рассмотренных подходов удовлетворяют требованиям, представленным во введении. Ни один из подходов не удовлетворяет всем требованиям. Для иерархических подходов выполняются максимум два критерия из четырех и максимум один из желаемых. В данной работе будут предложены два способа построения иерархии, один из которых будет удовлетворять всем четырем обязательным требованиям, а другой — трем обязательным требованиям и двум желаемым.

3 Послойное построение иерархии с помощью регуляризатора связи уровней

Как было отмечено в обзоре литературы, ARTM — это мощный инструмент для построения тематических моделей, позволяющий учитывать дополнительные критерия различного характера в одной модели. Это удобная основа для построения тематической иерархии, удовлетворяющей введенным требованиям. В этом подходе уже существует аппарат для разреживания тем, поэтому требование разреженности автоматически выполнено. Кроме того, разработана [17] онлайн-версия алгоритма обучения модели, позволяющая настраивать параметры, проходя по коллекции документов всего несколько раз. Если реализовать построение иерархии в ARTM, не усложняя алгоритм обучения одного уровня, то требование масштабируемости будет автоматически выполнено.

3.1 Обозначения и определения

При фиксированном уровне l иерархии будем обозначать S – множество тем данного уровня, T – множество родительских тем, то есть множеством тем $(l-1)$-го уровня. Матрицы родительского уровня будем обозначать $\Phi^p \in R^{|W|\times|T|}$ и $\Theta^p \in R^{|T|\times|D|}$.

Обозначим $\Psi = \{\psi_{st}\}_{s\in S, t\in T}$, $\psi_{st} = p(s|t)$, и будем называть ее матрицей перехода между уровнями. Величина ψ_{st} показывает вероятность перейти в подтему S из надтемы t.

Критерий	PLSA	LDA	ATM	FSTM	SWB	ARTM	hLDA	PAM	hPAM	hHDP	STROD	splitLDA
Иерархия	–	–	–	–	–	–	+	+	+	+	+	+
ТПО	·	·	·	·	·	·	+	–	+	–	–	+
ММ	–	–	+	–	–	+	–	–	–	–	–	–
СТ	·	·	·	·	·	·	–	+	+	+	–	–
ТР	–	–	–	+	–	+	–	–	–	–	–	–
ТМ	+	+	–	+	–	+	–	–	–	–	+	+
ТРД	·	·	·	·	+	·	+	–	+	–	–	–
ТРС	·	·	·	·	+	·	–	–	–	–	–	–

Таблица 1: *Сравнение тематических моделей. Точка в ячейке означает, что данный критерий не применим к неиерархической модели. Иерархия — является ли модель иерархической. ММ — поддерживает ли модель введение дополнительных модальностей, кроме текстовой. ТПО, СТ, ТР, ТМ, ТРД, ТРС — удовлетворяет ли модель требованию полноты описания, структурному требованию, требованию разреженности, требованию масштабируемости, требованиям расслоения документа и словаря. ТПО подразумевает, что модель позволяет описывать темы, как минимум, терминами и документами, СТ — модель поддерживает множественное наследование тем. На самом деле, это урезанное структурное требование; среди подходов, удовлетворяющих СТ, ни один не предоставляет возможностей для разреживания структуры графа. Требование масштабируемости считается невыполненным, если алгоритм обучения основан на семплировании Гиббса, в силу его недетерминированности [22]. Алгоритмы, для которых известны эффективные онлайн версии, считаются масштабируемыми. Требование расслоения словаря формально не выполнено ни для одной иерархической модели, потому что в них нет инструментов для разреживания дочерних тем. На практике, как правило, выполняется ослабленное ТРС, когда термины, оказавшиеся среди топовых в родительской теме, не показываются в списках топовых терминов дочерних. Однако проверить это свойство теоретически сложно, и оно не включено в сравнение.*

По аналогии будем называть матрицей обратного перехода между уровнями $\widetilde{\Psi} = \left(\tilde{\psi}_{ts}\right)_{t\in T, s\in S}$ $\tilde{\psi}_{ts} = p(t|s)$.

3.2 Регуляризатор матрицы терминов в темах

Будем строить иерархию сверху вниз, уровень за уровнем (каждый уровень – обычная плоская модель), при этом находя для каждой темы нового уровня родителей с предыдущего (уже построенного) уровня.

Будем приближать уже построенную матрицу Φ^p родительского уровня произведением $\Phi\Psi$:

$$p(w|t) \approx \sum_s p(w|s)p(s|t) = \sum_s \phi_{ws}\psi_{st}. \tag{5}$$

Данное разложение подразумевает введение еще одной *гипотезы условной независимости* $p(w|s,t) = p(w|s)$: вероятность появления термина w в подтеме s не зависит от надтемы t.

В качестве меры близости распределений возьмем дивергенцию Кульбака-Лейблера, и получим регуляризатор

$$R(\Phi) = \mathrm{KL}(\Phi^p||\Phi\ \Psi) = \sum_t \sum_w p(w|t) \ln \frac{p(w|t)}{\sum_s p(w|s)p(s|t)} = Const - \sum_t \sum_w \phi^p_{wt} \ln \sum_{s\in S} \phi_{ws}\psi_{st},$$

который необходимо минимизировать.

Новая оптимизационная задача:

$$\ln L(\Phi,\Theta,\Psi) = \sum_{d,w} n_{dw} \ln \sum_{s\in S} \phi_{ws}\theta_{sd} + \lambda \sum_{t\in T} \sum_{w\in W} \phi^p_{wt} \ln \sum_{s\in S} \phi_{ws}\psi_{st} + R(\Phi,\Theta,\Psi) \to \max_{\Phi,\Theta,\Psi}. \tag{6}$$

с ограничениями

$$\sum_w \phi_{ws} = 1; \phi_{ws} \geqslant 0, \quad \sum_s \theta_{sd} = 1; \theta_{sd} \geqslant 0, \quad \sum_s \psi_{st} = 1; \psi_{st} \geqslant 0. \tag{7}$$

При таком матричном разложении $\Phi^p \approx \Phi\Psi$ мы требуем, чтобы столбцы родительской Φ^p были линейной комбинацией столбцов дочерней Φ, то есть представляем родительскую тему как смесь дочерних тем.

Формула описанного регуляризатора имеет схожую структуру с формулой правдоподобия модели. Поэтому такая постановка задачи эквивалентна добавлению во входную матрицу n_{dw} $|T|$ псевдодокументов, отвечающих столбцам родительской Φ: $n_{d'w} = \lambda\phi_{wt}, d' = t$, и для обучения модели применим стандартный EM-алгоритм. Отвечающая введенным псевдодокументам часть Θ составит матрицу Ψ.

3.3 Регуляризатор матрицы тем в документах

Устанавливать взаимоотношения темы-подтемы можно и по матрице Θ. В этом случае матрица $\widetilde{\Psi}$ перехода между уровнями будет соответствовать вероятности надтемы t для темы s, а дополнительная задача матричного разложения и регуляризатор примут вид

$$p(t|d) \approx \sum_{s\in S} p(t|s)p(s|d) = \sum_{s\in S} \tilde{\psi}_{ts}\theta_{sd} \Leftrightarrow \Theta^p = \widetilde{\Psi}\Theta, \tag{8}$$

$$R(\Theta) = KL(\Theta^p||\Psi\Theta) = Const - \sum_{t\in T} \sum_{d\in D} \theta^p_{td} \ln \sum_{s\in S} \tilde{\psi}_{ts}\theta_{sd}.$$

Новая оптимизационная задача:

$$\sum_{d,w} n_{dw} \ln \sum_{s \in S} \phi_{ws}\theta_{sd} + \lambda \sum_{t \in T} \sum_{d \in D} \theta^p_{td} \ln \sum_{s \in S} \tilde{\psi}_{ts}\theta_{sd} + R(\Phi, \Theta, \tilde{\Psi}) \to \max_{\Phi,\Theta,\tilde{\Psi}}. \tag{9}$$

Эта модель в точности соответствует многомодальной версии APTM, если считать темы родительского уровня отдельной модальностью $\widetilde{W}$: $\tilde{w}_i = t_i$, $i = 1, \ldots, |T|$. Часть матрицы Φ, отвечающая данной модальности, составит $\widetilde{\Psi}$.

Отметим, что применять два введенных регуляризатора совместно не предполагается, потому что матрицы Ψ и $\widetilde{\Psi}$ связаны формулой Байеса, и независимая настройка матриц приведет к их рассогласованию.

Обратим также внимание, что предложенный метод не предоставляет единой генеративной модели коллекции: каждому уровню иерархии соответствует своя плоская модель, и эти модели связаны соотношениями (5) или (8). По этой причине два введенных регуляризатора были названы регуляризаторами связи уровней. Концептуально этот подход похож на модель hHDP, в котором дочерняя матрица Φ или перенормированная Θ подается на вход следующему уровню, но с движением сверху вниз и с добавлением родительских тем в коллекцию в качестве псевдодокументов или новой модальности.

Предложенный подход удовлетворяет требованиям полноты описания и структурному требованию. По аналогии с разреживанием Φ и Θ, можно ввести регуляризатор разреживания Ψ и $\widetilde{\Psi}$ и контролировать с его помощью степень разреженности графа. С точки зрения сложности обучения модели, мы добавили лишь малое число псевдодокументов или элементов модальности в коллекцию, что практически не влияет на скорость работы EM-алгоритма. Таким образом, предложенный подход удовлетворяет всем четырем требованиям, выдвинутым во введении.

Что касается двух дополнительных требований, то требованию расслоения документа подход не удовлетворяет. Требование расслоение словаря можно удовлетворить, если, согласно [25], разделять множество тем каждого уровня на фоновые и предметные и собирать в фоновых темах те термины, которые характерны для родительских тем. Однако в этом случае придется подбирать 4 дополнительных коэффициента регуляризации.

3.4 Эксперименты

В этом подразделе производится изучение влияния совместного использования регуляризаторов разреживания и введенных регуляризаторов связи уровней на модель, а также сравнение двух предложенных регуляризаторов.

Критерии качества. Автоматическое оценивание качества тематических иерархий остается открытой научной проблемой [28]. Введем несколько критериев для оценивания мо-

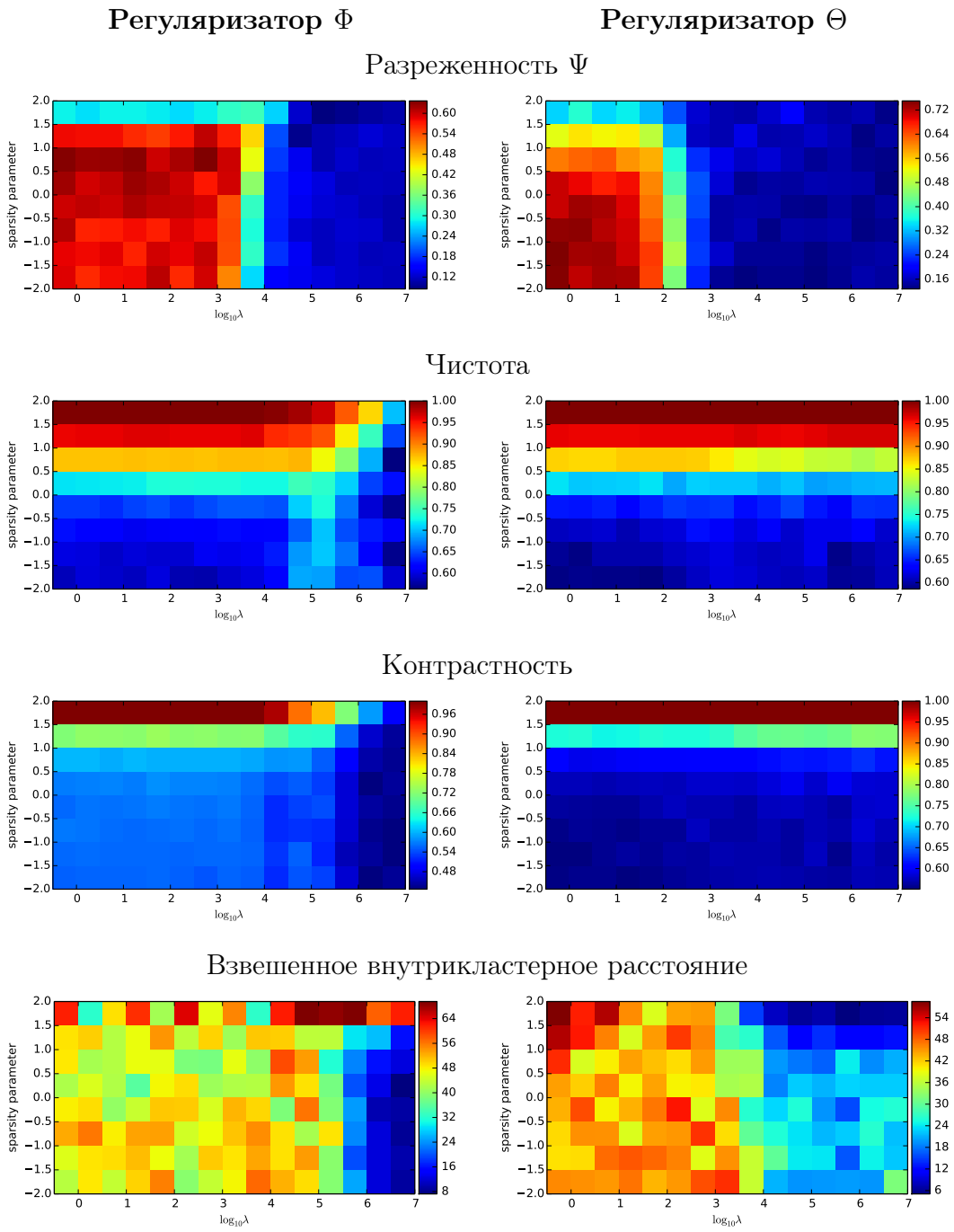


Рис. 1: *Исследование взаимодействия сглаживающих иерархических регуляризаторов (лог. шкала по оси абсцисс) и разреживающих регуляризаторов повышения интерпретируемости (лог. шкала по оси ординат). Рассматриваются матрицы перехода* Ψ *и* $\widetilde{\Psi}$ *между первым и вторым уровнем иерархии,* $|S| = 30$*. Все величины усреднены по 5 запускам алгоритма обучения.*

дели, отвечающие за различные аспекты качества.

Для оценки интепретируемости иерархии используются критерии, предложенные в [25].

В каждой теме выделяется *лексическое ядро* — множество $\mathrm{K}(s)$ слов w, для которых $p(s|w) > 0.25$ — и оцениваются его параметры: количество слов *S(s)*, *чистота*

$$\mathrm{P}(s) = \sum_{w \in \mathrm{K}(s)} p(w|s).$$

и *контрастность*

$$C(s) = S(s)^{-1} \sum_{w \in K(s)} p(s|w).$$

Чтобы сравнивать качество кластеризации документов по темам одного уровня, предлагается вычислять математическое ожидание похожести двух различных случайно вытащенных из темы документов. Чем оно больше, тем более вероятно, что человек сочтет документы, отнесенные к теме, близкими по смыслу. Документ может принадлежать нескольким темам, поэтому элементы, входящие в него, разделяются между темами в соответствии с формулами Е-шага: $n_d^s = \{n_{dw}^s\}_{w \in W}, n_{dw}^s = n_{dw} p(s|d,w)$. Критерий учитывает пары наиболее вероятных для темы документов с большим весом:

$$I_s = \sum_{d_1, d_2 : \theta_{s,d1} \neq 0, \theta_{s,d2} \neq 0} \mathrm{sim}(n_{d_1}^s, n_{d_2}^s) p(d_1, d_2|s),$$

$$\mathrm{sim}(n_{d_1}, n_{d_2}) = \sum_{w \in W} n_{d_1 w} n_{d_2 w}, \quad p(d_1, d_2|s) \propto p(d_1|s) p(d_2|s), d_1 \neq d_2.$$

Критерием плотности иерархического графа является плотность матрицы Ψ — среднее число тематических связей между уровнями:

$$G = \frac{1}{|T||S|} \sum_t \sum_s [\psi_{ts} > \epsilon].$$

На практике оказывается, что большинство значений матрицы Ψ хоть и не обнулятся, но становятся очень малыми, и их не нужно учитывать при построении навигатора. В экспериментах $\epsilon = 0.05$.

Для контроля качества дополнительного матричного разложения выбрано расстояние Хеллингера между матрицами $A = \{a\}_{i,j=1}^{m,n}$ и $B = \{b\}_{i,j=1}^{m,n}$ одного размера:

$$\rho(A,B) = \frac{1}{m} \sum_{j=1}^{m} \sqrt{\frac{1}{2} \sum_{i=1}^{n} (\sqrt{a_{ij}} - \sqrt{b_{ij}})^2}.$$

Коллекция документов. Для проведения экспериментов подготовлена лемматизированная и преобразованная в матрицу частот слов коллекция материалов конференций *Математические методы распознавания образов* и *Интеллектуализация обработки информации*[1] за 2007 —2013 года. В коллекции 850 документов, размер словаря — 50856 словосочетаний. Словосочетания выделены с использованием внешних средств.

[1] http://mmro.ru

На этой коллекции строится трехуровневая иерархия с количеством тем на уровнях 10, 30, 60.

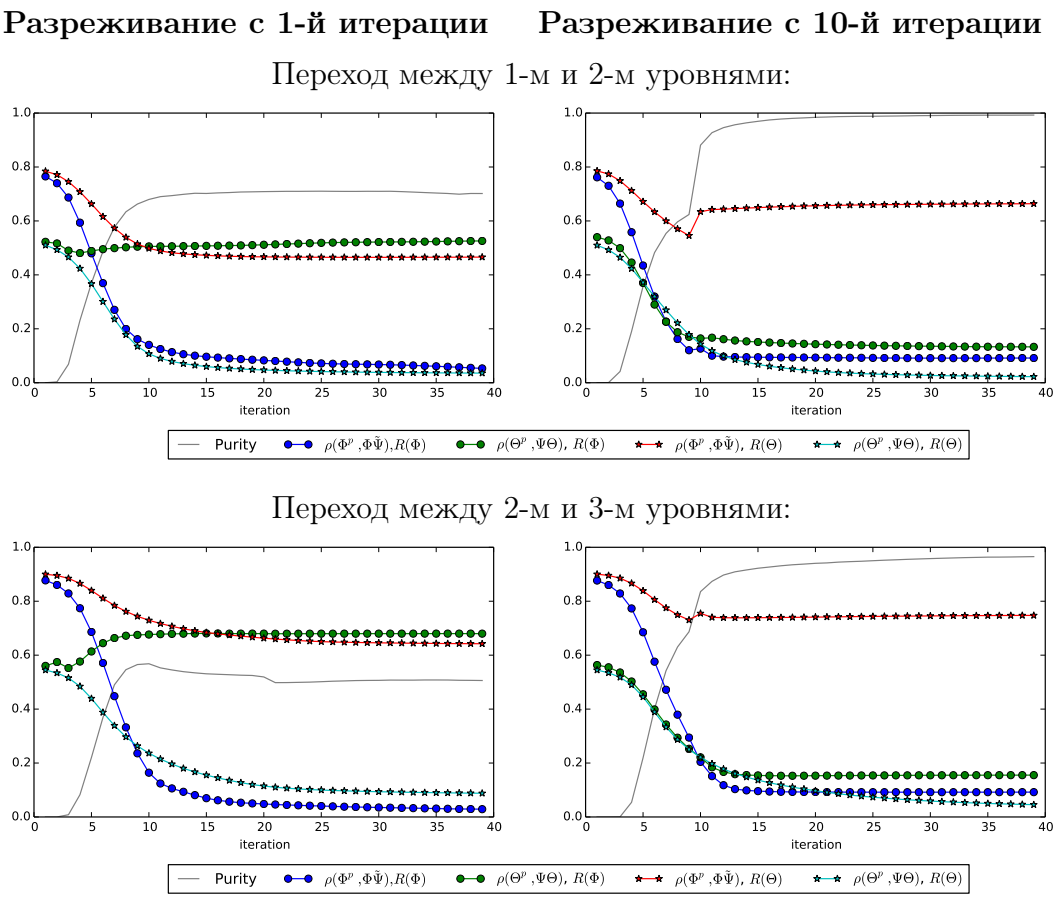


Рис. 2: *Исследование качества аппроксимации родительских тем смесью дочерних. По оси абсцисс — номер итерации EM-алгоритма, по оси ординат — расстояние между исходными матрицами родительского уровня* Φ^p *и* Θ^p *и их стохастическими разложениями* $\Phi\Psi$ *и* $\widetilde{\Psi}\Theta$ *на дочернем уровне. Переход от* Ψ *к* $\widetilde{\Psi}$ *осуществляется по формуле Байеса:* $p(t|s) = \frac{p(s|t)p(t)}{\sum_t p(s|t)p(t)}$. $R(\Phi)$ *и* $R(\Theta)$ *указывают, какой регуляризатор применялся при построении модели, серым цветом показано значение чистоты модели. В моделях, соответствующих графикам правой колонки, разреживание включалось с 10-й итерации; так удается достичь лучших показателей интерпретируемости. Коэффициенты регуляризации одинаковые для моделей, соответствующих одному переходу, и подобраны в ходе предыдущего эксперимента.*

Изучение влияния регуляризаторов на модель. Чтобы проанализировать, как сглаживающий иерархический регуляризатор влияет на качество модели в сочетании с разреживающими регуляризаторами, проведена серия экспериментов с перебором параметров обоих регуляризаторов по двумерной сетке (рис. 1).

Для обоих иерархических регуляризаторов существует значение параметра, при котором Ψ получается достаточно разреженной (а многодольный граф приближается к дереву). Параметр регуляризатора Θ практически не влияет на интерпретируемость модели, но зато ухудшает показатель качества кластеризации. Регуляризатор Φ при некотором значении начинает ухудшать чистоту и контрастность, однако в сочетании с высокой разреженностью модели он устойчиво дает лучшую кластеризацию документов (по критерию I), чем иные комбинации, при приемлемом значении остальных критериев.

Сравнение двух предложенных регуляризаторов. Чтобы сравнить свойства двух предложенных регуляризаторов, исследуется качество аппроксимации родительских тем смесью дочерних (рис. 2). Рассматривается расстояние от родительских матриц Φ и Θ до их аппроксимации на дочернем уровне и его зависимость от номера итерации EM-алгоритма. Эксперимент проводился для двух стратегий регуляризации (когда разреживание Φ включалось с 1-й и с 10-й итерации) и для двух переходов между уровнями.

Для всех восьми случаях матрица, на которую действовал регуляризатор, аппроксимируется с достаточной точностью. Однако регуляризатор Φ при использовании рекомендуемой стратегии разреживания аппроксимирует еще и матрицу Θ^p, то есть выполняет функцию регуляризатора Θ.

На основании проведенных экспериментов для построения финальной иерархической модели по коллекции ММРО-ИОИ был выбран регуляризатор матрицы Φ.

Составление иерархического тематического навигатора на основе модели. Как упоминалось во введении, основной целью построения иерархических тематических моделей является организация удобного доступа к документам и навигации по коллекции. Разработана визуализация описанной иерархической модели, предполагающая выделение отдельных страниц для тем и документов, а также отображение структуры иерархии в виде графа. При нажатии на вершину в графе открывается панель с ее описанием и ссылкой на отвечающую ей страницу; для тем на панели также отображаются ранжированные по $p(w|s)$ и $p(d|s)$ топы терминов и документов этой темы.

Визуализация доступна по адресу *explore-mmro.ru*.

3.5 Выводы

Сформулируем основные выводы раздела:

1. Введение регуляризатора связи тем позволяет построить иерархическую тематическую модель, удовлетворяющую четырем требованиям, изложенным во введении.

Эксперименты подтверждают, что при качественном подборе коэффициентов и траектории регуляризации тематические распределения над терминами в построенной модели будут разреженными, а число ребер в графе будет маленьким. При этом на разреженность графа сильно влияет коэффициент при регуляризаторе связи уровней.

2. В равных условиях регуляризатор матрицы терминов в темах позволяет аппроксимировать обе родительских матрицы, в то время как регуляризатор матрицы тем в документах позволяет приблизить только родительскую матрицу Θ.

3. Начинать разреживание параметров модели лучше не с первой итерации, а когда модель уже достаточно хорошо сошлась. Этот вывод согласуется с результатами [25].

4 Единая иерархическая модель коллекции

Подход, предложенный в предыдущем разделе, удовлетворяет выдвинутым требованиям, но для построения такой иерархии требуется подбирать много коэффициентов регуляризации. При этом идея регуляризатора Θ выглядит логично: множество документов, составляющих родительскую тему, является объединением документов дочерних тем. Введение регуляризатора Φ подразумевает принятие такой же гипотезы, но относительно множества терминов родительской темы, и это предположение уже противоречит требованию расслоения словаря. В то же время, если принять первую гипотезу, можно построить модель коллекции, которая будет удовлетворять обоим требованиям расслоения. Об этом пойдет речь в данном разделе.

4.1 Обозначения и определения

Пусть иерархия — это многодольный граф тем с уровнями $1, \dots, L$ и множествами тем $S^1, \dots, S^L$ соответственно, $|S^L| > |S^{L-1}| > \cdots > |S^1|$. В том числе, может быть $|S^1| = 1$.

Введем следующие матрицы параметров:

- $\Phi^1, \dots, \Phi^L$: $\phi^l_{ws^l} = p(w|s^l), w \in W, s^l \in S^l, l = 1, \dots, L$;
- $\Psi^1, \dots, \Psi^{L-1}$: $\psi^l_{s^l, s^{l+1}} = p(s^l|s^{l+1}), s^l \in S^l, s^{l+1} \in S^{l+1}, l = 1, \dots, L-1$;
- Θ: $\theta_{s^L d} = p(s^L|d), s^L \in S^L, d \in D$;
- H: $\eta_{ld} = p(l|d), l = 1, \dots, L, d \in D$.

Первая группа параметров — это матрицы распределений терминов в темах для каждого уровня, вторая — матрицы переходов между уровнями, то есть переходов с нижнего уровня на верхний, третья — матрица распределений тем последнего уровня в документах, и четвертая — матрица распределений уровней в документах.

4.2 Описание и обучение модели

Будем считать, что генерация слова w_i в документе d, $i = 1, \ldots, n_d$ происходит согласно следующей модели:

1: Семплируется тема последнего уровня $s^L \sim p(s^L|d) = \text{Mult}(\theta_d)$;
2: Семплируется номер уровня $l \sim p(l|d) = \text{Mult}(\eta_d)$;
3: **для** $l' = L, \ldots, l+1$:
4: семплируется надтема $s^{l'-1} \sim p(s^{l'-1}|s^{l'}) = \text{Mult}(\psi_{s^{l'}}^{l'-1})$;
5: из выбранной темы s^l уровня l семплируется термин $w \sim p(w|s^l) = \text{Mult}(\phi_{s^l}^l)$.

Будем называть последнюю выбранную тему s^l уровня l *финальной*.

Как и в hLDA и hPAM, термины в модели могут семплироваться со всех уровней. Как в hLDA, у каждого документа есть распределение над уровнями. Как в hHDP версии с документами, документы описываются распределениями только над темами нижнего уровня, при этом

$$p(s^l|d) = p(s^l|s^{l+1}) \ldots p(s^{L-1}|s^L)p(s^L|d).$$

При этом в модели не накладываются априорные распределения на параметры и по аналогии с PLSA в процессе обучения необходимо настраивать сами матрицы параметров.

Распределение $p(l|d)$ показывает, какая часть терминов документа относится к темам l-го уровня, то есть оценивает, насколько узка тематика документа, это означает выполнение требования расслоения документа. Как и в других иерархических моделях, при создании навигатора в темах можно выводить термины и документы, отсортированные по убыванию вероятности их появления в теме. Но данная модель более гибкая: если сортировать документы по убыванию $p(d|s^l) \propto p(s^l|d)p(d)$, то среди топовых мы увидим самые важные документы темы, а если по $p(s^l, l|d) \propto p(s^l|d)p(l|d)p(d)$, то мы увидим документы с наиболее общей тематикой в рамках данной темы, а более узкие по тематике документы окажутся топовыми в списках документов дочерних тем. Как и другие иерархические модели, данная модель удовлетворяет ослабленному условию расслоения словаря: топовые термины родительских тем редко встречаются среди топовых терминов дочерних. Ниже будет показано, что с помощью специального метода разреживания в рамках подхода можно удовлетворить и само требование расслоения словаря.

Таким образом, предложенная модель предоставляет более удобный интерфейс для создания тематического навигатора, чем другие подходы.

Обучение модели. Запишем совместное распределение на наблюдаемые переменные и параметры модели и маргинализуем его по параметрам:

$$p(w,l,s^l,\dots,s^L|d) = p(w|s^l)p(s^l|s^{l+1})\dots p(s^{L-1}|s^L)p(s^L|d)p(l|d) = \phi^l_{ws^l}\psi^l_{s^l,s^{l+1}}\dots\psi^{L-1}_{s^{L-1},s^L}\theta_{s^Ld}\eta_{ld} \tag{10}$$

$$p(w|d) = \sum_{l,s^l,\dots,s^L} p(w,l,s^l,\dots,s^L|d); \tag{11}$$

Здесь подразумевается ряд гипотез условной независимости, в частности, выбор термина в финальной теме и выбор надтем не зависят от документа и предыдущих выбранных тем.

Как и в APTM, предлагается настраивать параметры методом максимизации регуляризованного логарифма правдоподобия:

$$\ln L(\Phi^1,\dots,\Phi^L,\Psi^1,\dots,\Psi^{L-1},\Theta,H) = \sum_{d\in D}\sum_{w\in d} n_{dw}\ln p(w|d) =$$

$$= \sum_{d\in D}\sum_{w\in d} n_{dw}\ln\sum_{l,s^l,\dots,s^L}\phi^l_{ws^l}\psi^l_{s^l,s^{l+1}}\dots\psi^{L-1}_{s^{L-1},s^L}\theta_{s^Ld}\eta_{ld}; \tag{12}$$

$$\ln L + R(\Phi^1,\dots,\Phi^L,\Psi^1,\dots,\Psi^{L-1},\Theta,H) \to \max_{\Phi^1,\dots,\Phi^L,\Psi^1,\dots,\Psi^{L-1},\Theta,H}. \tag{13}$$

$$\phi^l_{ws^l}\geqslant 0, \sum_w \phi^l_{ws^l} = 1;\ \psi^l_{s^l,s^{l+1}}\geqslant 0, \sum_{s^l}\psi^l_{s^l,s^{l+1}} = 1;\ \theta_{s^Ld}\geqslant 0, \sum_{s^l}\theta_{s^Ld} = 1;\ \eta_{ld}\geqslant 0, \sum_l \eta_{ld} = 1. \tag{14}$$

Теорема 2 *Если $R(\Phi^1,\dots,\Phi^L,\Psi^1,\dots,\Psi^{L-1},\Theta,H)$ непрерывно дифференцируема по своим параметрам, то стационарная точка задачи* (4.2)—(14) *удовлетворяет следующей системе уравнений:*

$$\phi^l_{ws^l} = \underset{w}{\operatorname{norm}}\frac{\partial(\ln L + R)}{\partial\phi^l_{ws^l}}\phi^l_{ws^l}, \qquad w\in W, s^l\in S^l, l = 1,\dots,L; \tag{15}$$

$$\psi^l_{s^l,s^{l+1}} = \underset{s^l}{\operatorname{norm}}\frac{\partial(\ln L + R)}{\psi^l_{s^l,s^{l+1}}}\psi^l_{s^l,s^{l+1}}, \quad s^l\in S^l, s^{l+1}\in S^{l+1}, l = 1,\dots,L-1; \tag{16}$$

$$\theta_{s^Ld} = \underset{s^L}{\operatorname{norm}}\frac{\partial(\ln L + R)}{\partial\theta_{s^Ld}}\theta_{s^Ld}, \qquad s^L\in S^L, d\in D \tag{17}$$

$$\eta_{ld} = \underset{l}{\operatorname{norm}}\frac{\partial(\ln L + R)}{\partial\eta_{ld}}\eta_{ld}, \qquad l = 1,\dots,L, d\in D. \tag{18}$$

Здесь и далее будем опускать параметры логарифма правдоподобия и регуляризаторов.

Доказательство. Воспользуемся теоремой Каруша-Куна-Такера. Проигнорируем требования неотрицательности и учтем их позже.

Введем двойственные переменные $\xi = \{\xi_{l,s^l}\}_{s^l \in S^l, l=1,\dots,L}$, $\zeta = \{\zeta_{l,s^l}\}_{s^l \in S^l, l=1,\dots,L-1}$, $\alpha = \{\alpha_d\}_{d \in D}$, $\beta = \{\beta_d\}_{d \in D}$. Запишем лагранжиан:

$$\mathscr{L}(\Phi^1, \dots, \Phi^L, \Psi^1, \dots, \Psi^{L-1}, \Theta, H; \xi, \zeta, \alpha, \beta)$$

$$= \ln L + R - \sum_l \sum_{s^l} \xi_{ls^l} \Big(\sum_w \phi^l_{ws^l} - 1 \Big) - \sum_l \sum_{s^l} \zeta_{ls^l} \Big(\sum_{\hat{l}} \psi^l_{s^{\hat{l}} s^l} - 1 \Big) - \sum_d \alpha_d \Big(\sum_{s^L} \theta_{s^L d} - 1 \Big) - \sum_d \beta_d \Big(\sum_l \eta_{ld} - 1 \Big).$$

Продифференцируем лагранжиан по параметрам модели, приравняем к нулю и домножим на переменные дифференцирования:

$$\frac{\partial(\ln L + R)}{\partial \phi^l_{ws^l}} \phi^l_{ws^l} - \xi_{ls^l} \phi^l_{ws^l} = 0; \tag{19}$$

$$\frac{\partial(\ln L + R)}{\partial \psi^l_{s^l, s^{l+1}}} \psi^l_{s^l, s^{l+1}} - \zeta_{ls^l} \psi^l_{s^l, s^{l+1}} = 0; \tag{20}$$

$$\frac{\partial(\ln L + R)}{\partial \theta_{s^L d}} \theta_{s^L d} - \alpha_d \theta_{s^L d} = 0; \tag{21}$$

$$\frac{\partial(\ln L + R)}{\partial \eta_{ld}} \eta_{ld} - \beta_d \eta_{ld} = 0. \tag{22}$$

Перенесем произведения с двойственными переменными в (19)–(22) в правую часть. Поскольку правая часть всех получающихся выражений неотрицательна, заменим левые части выражений на их положительные срезки: $(z)_+ = \max\{z, 0\}$; просуммируем каждую из групп выражений по первому индексу. Получим, что двойственные переменные равны нормировочной константе своих распределений. В итоге придем к выражениям (15)—(18).

Для обучения модели к системе (15)—(18) применяется метод простой итерации, то есть значения параметров на новой итерации обновляются по формулам, представленным в правых частях уравнений системы. Изначально параметры модели инициализируются случайными числами и нормируются. Вычисление дифференциалов $\ln L$ не составляет технических сложностей, поскольку матричное разложение линейно.

Эвристический способ разреживания параметров модели В отличие от других генеративных иерархических моделей, в предложенной не накладываются априорные распределения на параметры. Это позволяет воздействовать напрямую на распределения, участвующие в модели, чтобы они удовлетворяли дополнительным критериям.

Предположим, мы хотим, чтобы некоторое распределение $\alpha = (\alpha_1, \dots, \alpha_K), \alpha_i \geqslant 0, \sum_{i=1}^K \alpha_i = 1$ было разреженным. Будем решать следующую оптимизационную задачу:

$$\frac{1}{p} \sum_{i=1}^K \alpha_i^p \to \max_\alpha,$$

где $p > 1$. Легко показать, что максимум этого распределения достигается при $\alpha_i = 1, \alpha_j = 0, j \neq i$. Применив теорему Каруша-Куна-Такера, получим:

$$\alpha_i^{p-1} - \xi = 0,$$

ξ — двойственная переменная. Тогда, домножив на α_i, получим

$$\xi = \sum_i \alpha_i^p, \alpha_i^{new} = \text{reg}(\alpha_i) = \frac{\alpha_i^p}{\sum_j \alpha_j^p}.$$

Теоретически ни одна компонента вектора α не будет равна нулю, но на практике это произойдет из-за ограниченности машинной точности.

Оператор reg будет применяться к разреживаемому распределению после каждой итерации основного алгоритма обучения модели.

Очевидно, что применение оператора reg с параметром p n раз равносильно единичному его применению с параметром p^n, то есть p выполняет функцию балансирования между максимизацией правдоподобия и разреживанием. В то же время, величина p не зависит от масштаба данных, и его можно применять со стандартным значением $p = 2$.

Однако важнее другое свойство: при описанном подходе к разреживанию мы не можем полностью обнулить распределение, как это происходит в APTM, если задать большой коэффициент при регуляризаторе.

Хотя такой подход не имеет никаких теоретических гарантий сходимости алгоритма, в экспериментах показывается, что после нескольких итераций с начала разреживания правдоподобие модели стабилизируется.

Несколько примеров конкретных регуляризаторов для иерархической модели:

- разреживание $p(s, l|w)$: при таком разреживании мы получаем чистоту темы, очень близкую к единице, это означате, что почти все термины монотематичны. Для применения такого разреживания необходимо перейти от $\phi_{ws^l} = p(w|s^l, l)$ к $p(s^l, l|w)$ с помощью формулы Байеса, применить оператор reg и перейти обратно к переменным ϕ_{ws^l}.

- разреживание $p(s^l|s^{l+1}) = \psi_{s^l,s^{l+1}}$: в этом случае мы стремимся сделать иерархию похожей на дерево. Здесь важно указанное свойство о том, что распределение не может полностью обнулиться. Иначе граф мог бы стать несвязным.

Анализ модели Обозначим $\Lambda^l = diag\{\eta_{ld_1}, \ldots, \eta_{ld_{|D|}}\}$. Выражения (10)–(11) равносильны следующему линейному разложению матрицы $F = \{f_{dw}\} = \{\frac{n_{dw}}{n_d}\}$:

$$F \approx \Phi^L \Theta \Lambda^L + \Phi^{L-1}\Psi^{L-1}\Theta\Lambda^{L-1} + \cdots + \Phi^1\Psi^1 \ldots \Psi^{L-1}\Theta\Lambda^1.$$

Это разложение по уровням можно воспринимать как взвешенную сумму обычных плоских тематических моделей с новым обозначением

$$\Theta^l = \Psi^l \dots \Psi^{L-1}\Theta; \tag{23}$$

$$F \approx \Phi^L\Theta^L\Lambda^L + \Phi^{L-1}\Theta^{L-1}\Lambda^{L-1} + \dots + \Phi^1\Theta^1\Lambda^1.$$

Получается, что мы находим представление матрицы частот слов в виде суммы нескольких глубоких матричных разложений.

Простейшим примером модели служит топология «корень-K тем». Тогда корневая тема будет фоновой, все вероятности перехода между уровнями $p(t|s) = 1$, а $\eta_d = \{\eta_{0,d}, \eta_{1,d}\}$ показывает соотношение фоновой и предметной лексики в документе. Эта модель очень похожа на ту, которая была предложена в SWB [8].

Благодаря применению описанного способа разреживания удается добиться того, чтобы матрицы Φ^l имели ненулевые вероятности терминов только своего уровня, а не всех дочерних уровней, то есть чтобы модель удовлетворяла требованию расслоения словаря.

Сложность обучения модели линейна по количеству документов, хотя, как правило, требует достаточно много проходов по коллекции до сходимости. Стохастические методы оптимизации для этой пока задачи не пробовались.

Модель допускает множественное наследование тем, при этом предложен способ уменьшения числа ребер в нем. Темы описываются терминами и документами.

Таким образом, модель удовлетворяет всем требованиям, за исключением, быть может, требования масштабируемости, потому что применимость более быстрых алгоритмов оптимизации для обучения модели не исследована.

4.3 Эксперименты

В этом подразделе исследуется влияние разреживания на качество модели.

Критерии качества Сходимость алгоритма обучения модели оценивается по логарифму правдоподобия.

Плотность графа вычисляется как доля ненулевых элементов во всех матрицах Ψ^l.

Для оценки интерпретируемости тем используется когерентность — популярная мера качества тематических моделей. Считается, что показатель когерентность хорошо коррелирует с экспертными оценками интерпретируемости тем. В литературе когерентность определяют по-разному, мы будем пользоваться определением из [16]. Пусть $w = (w_1, \dots, w_{10})$ — 10 топовых терминов темы s. Тогда когерентность темы определяется как

$$c(w) = \frac{1}{45}\sum_{i=1}^{10}\sum_{j>i} PMI(w_i, w_j), \quad , PMI(w_i, w_j) = \ln\frac{p(w_i, w_j) + \epsilon}{p(w_i)p(w_j)},$$

величины $p(w_i, w_j)$, $p(w_i)$ и $p(w_j)$ показывают вероятности появления терминов в коллекции. Их можно оценивать по той же коллекции, а можно — по внешней, например, по Википедии. Мы будем оценивать эти вероятности по коллекции, на которой обучалась модель: $p(w_i, w_j)$ пропорционально количеству раз, когда термины w_i, w_j встретились вместе в окне длины 10. Параметр $\epsilon = 10^{-6}$ нужен для предотвращения взятия логарифма нуля. Когерентность модели будет равна средней когерентности по всем темам иерархии, включая корневую, если она есть. Чем больше когерентность, тем лучше.

Для оценивания качества кластеризации документов по темам модифицируем критерий однородности кластера [21]. Критерий однородности качества кластеризации вводится следующим образом: пусть алгоритм нашел в данных кластеры $c \in C$, при этом имеется эталонная разметка объектов на классы $k \in K$, $n = |C|$; обозначим a_{ck} — количество объектов, отнесенных к кластеру c и принадлежащих классу k. Введем величины

$$H(C|K) = -\sum_{c \in C} \sum_{k \in K} \frac{a_{ck}}{n} \ln \frac{a_{ck}}{\sum_{c \in C} a_{ck}}$$

$$H(C) = -\sum_{c \in C} \frac{\sum_{k \in K} a_{ck}}{n} \ln \frac{\sum_{k \in K} a_{ck}}{n}.$$

Тогда однородность кластеризации оценивается как

$$h = \begin{cases} 1, H(C|K) = 0 \\ 1 - \frac{H(C|K)}{HC}, H(C|K) \neq 0 \end{cases}$$

Если все кластеры представлены элементами только одного класса, однородность будет равна 1, чем больше классов представлено в кластере, тем ближе однородность к 0. Именно такую величину мы хотим оценивать для тем одного уровня: документы темы должны по возможности принадлежать одной смысловой категории.

Предположим, что для каждого документа коллекции известен класс, которому он принадлежит. Это свойство может выполняться только для простых коллекций, в которых каждый документ монотематичен. Тематическая модель все равно будет иметь возможность отнести документ к нескольким темам. Обозначим $a_{ck} = \sum_{d \in k} \theta_{cd}$ — обобщение обозначения a_{ck} на случай мягкой кластеризации, то есть такой, когда документ может относиться к нескольким темам-кластерам. Тогда по формулам, приведенным выше, можно оценить однородность темы. Однородность уровня модели будет определяться как средняя однородность тем этого уровня. Матрицы Θ для промежуточных уровней иерархии можно вычислить по формуле (23).

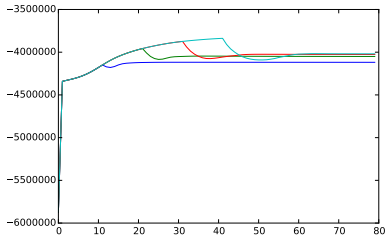
Коллекция документов. Эксперименты проводились на коллекции школьных конспектов, скачанных с сайта, помогающего школьникам готовиться к ЕГЭ. В коллекции

1491 документ, 27263 слов. Коллекция лемматизирована и приведена к виду матрицы частот слов.

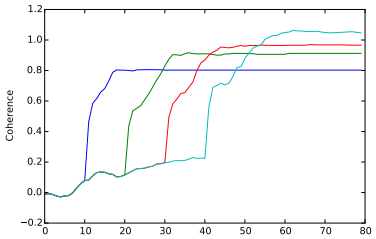
Во всех экспериментах строится трехуровневая иерархия с количеством тем на уровнях 1, 10, 30, регуляризация не применяется.

Для каждого документа известен класс — школьный предмет, к которому относится конспект. Эта информация используется для оценки однородности тем.

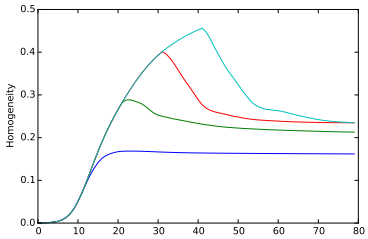
Изучение влияния разреживания матриц терминов в темах на качество модели. На качество модели оказывают влияние два параметра разрежвания — коэффициент p и номер итерации, с которого модель начинают разреживать. Проведены две серии экспериментов. В первой серии (рис.3) параметр фиксирован $p = 2$, а разреживание модели начинается с 10, 20, 30, 40 итерации. Все модели строились из одинакового начального приближения.



$\ln L$



Когерентность



Однородность посл. уровня

Рис. 3: *Исследование влияния стартовой итерации разреживания на качество модели. По оси абсцисс — номер итерации алгоритма обучения, по оси ординат — метрика качества. Синяя, зеленая, красная и голубая линии соответствуют разреживанию с 10, 20, 30 и 40 итераций соответственно. В модели 3 уровня, 1/10/30 тем. На плотность графа регуляризатор не влияет, поэтому графики для этого критерия не приводятся.*

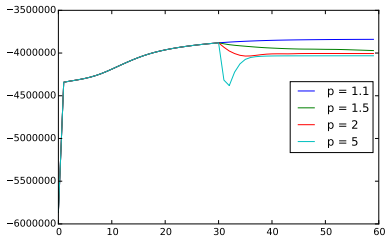
Разреживание уменьшает правдоподобие, но через несколько итераций оно вновь начинает расти. Итоговое правдоподобие разреженной модели ниже правдоподобия простой модели, однако чем позже начинается разреживание, тем выше в итоге поднимается $\ln L$. Аналогично происходит с когерентностью: при более позднем разреживании итоговая модель получается более интерпретируемой. Однородность сильно падает из-за разреживания, хотя также чуть-чуть увеличивается при более позднем разреживании.

Приведем примеры нескольких тем последнего уровня для разных значений стартовых итераций разреживания, жирным шрифтом выделены фоновые, часто встречающиеся во всей коллекции слова:

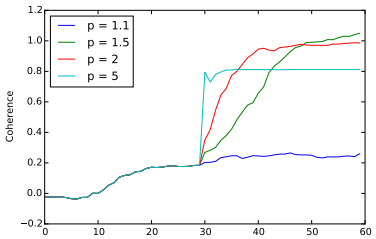
10	**при**, энергия, изменение, магнитный, заряд, волна, **свойство**, линия, расстояние	дело, император, идея, орган, **там**, **поздно**, николай, философский, **читать**	a, x, b, **следующий**, формула, x, прямой, f, ноль
20	**при**, **можно**, энергия, **через**, направление, поле, магнитный, заряд, волна	дело, **где**, император, смерть, реформа, **хороший**, сын, орган, **там**	a, x, b, **получать**, пример, уравнение, x, равный, **следующий**
30	энергия, направление, поле, магнитный, заряд, электрический, волна, линия, проводник	царь, император, смерть, александр, реформа, сын, восстание, **там**, дело	a, x, b, равный, выражение, формула, дробь, корень, ноль
40	поле, магнитный, заряд, электрический, волна, направление, линия, проводник, электромагнитный	царь, император, смерть, александр, реформа, церковь, правление, сын, восстание	a, b, выражение, дробь, корень, sin, степень, формула, cos

В темах последних двух моделей фоновые слова появляются значительно реже.

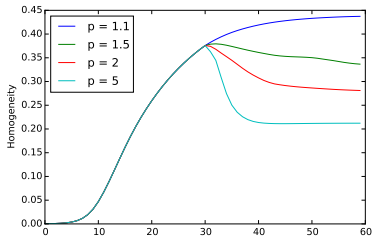
Во второй серии экспериментов (рис. 6) фиксирована стартовая итерация разрежи-вания, равная 30, а параметр $p \in \{1.1, 1.5, 2, 5\}$. Все модели строились из одинакового начального приближения.



$\ln L$



Когерентность



Однородность посл. уровня

Рис. 4: *Исследование влияния коэффициента p разреживания на качество модели. По оси абсцисс — номер итерации алгоритма обучения, по оси ординат — метрика качества. На плотность графа регуляризатор не влияет, поэтому графики для этого критерия не приводятся.*

Чем больше коэффициент разреживания, тем ниже итоговое значение правдоподобия и однородности тем. С когерентностью зависимость не монотонная: при $p = 1.5$ достигается оптимум когерентности, причем если для очень большого и очень маленького значения параметра когерентность стабилизируется, то для оптимального значения она продолжает расти.

Выведем таблицу тем. Модели первой и второй серий экспериментов строились из разных начальных приближений, и темы в них, вообще говоря, разные. Для демонстрации найдены похожие темы на те, которые были выбраны в первой серии:

1.1	поле, магнитный, **в**, ток, **быть**, линия, электромагнитный, свет, индукция	**в**, **и**, империя, **на**, **который**, **быть**, **с**, **свой**, россия	число, **на**, **и**, **в**, **который**, **быть**, **например**, **не**, цифра
1.5	ток, поле, магнитный, свет, линия, проводник, электромагнитный, индукция, поток	империя, российский, император, россия, правление, внешний, фактически, александр, i	число, единица, делиться, обозначать, натуральный, простой, записывать, деление, цифра
2	ток, поле, магнитный, свет, линия, проводник, электромагнитный, поток, индукция	российский, империя, внешний, император, начинаться, александр, правление, фактически, вступать	число, **например**, единица, обозначать, стоять, делиться, **д**, класс, натуральный
5	поле, магнитный, **вызывать**, свет, постоянный, линия, проводник, открывать, **вокруг**	**после**, **со**, **смочь**, российский, империя, внешний, император, борьба, великий	число, **например**, **без**, название, класс, единица, обозначать, делиться, **д**

Таким образом, $p = 1.5$ можно считать оптимальным.

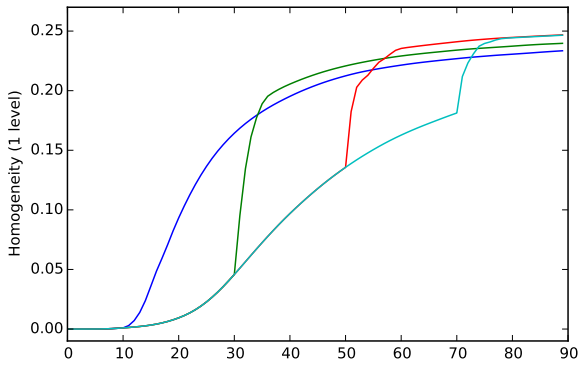
Исследование влияния разреживания графа на качество модели Аналогичные эксперименты были проведены с разреживанием матрицы Ψ^1. В третьей серии экспериментов разреживание начиналось с 10, 30, 50 и 70 итераций. Все модели строились из одного начального приближения. Разреживание распределений над терминами не производилось.

При любой стартовой итерации разреживания графа он в какой-то момент становился деревом. Так происходит из-за того, что в коллекции все документы монотематичны. Большие номера стартовых итераций увеличивают однородность тем промежуточного уровня, но не сильно.

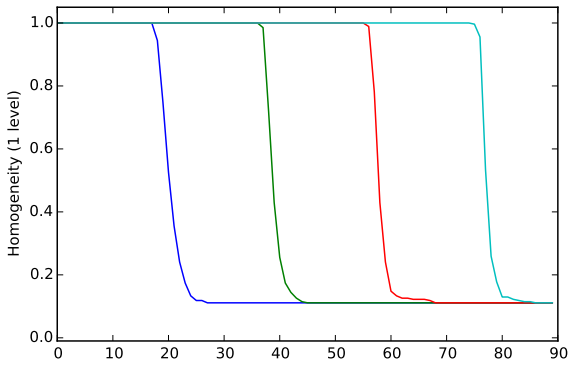
Можно сделать вывод, что разреживание тематического графа можно начинать с 50 итерации для данной коллекции.

В четвертой серии экспериментов перебираются те же значения коэффициента разреживания, что и во второй.

Для любого, даже очень маленького, коэффициента разреживания существует номер итерации, с которого граф принимает вид дерева. Для однородности это утверждение не верно: этот критерий меньше для $p = 1.1$ и одинаковый для других p.

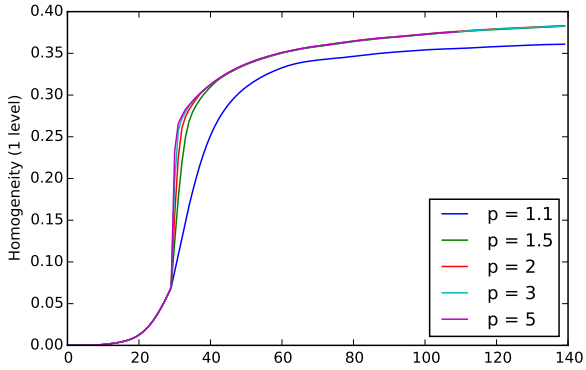


Однородность промежуточного уровня

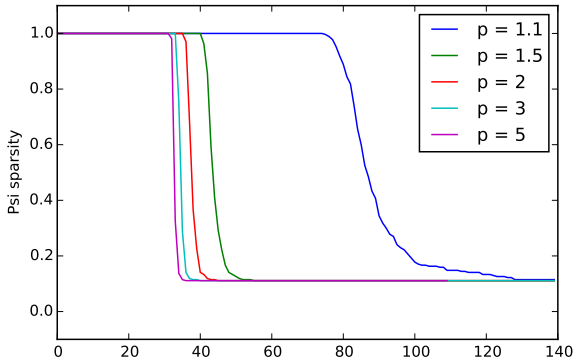


Плотность Ψ^1

Рис. 5: *Исследование влияния стартовой итерации разреживания графа на качество модели. По оси абсцисс — номер итерации алгоритма обучения, по оси ординат — метрика качества. На правдоподобие, когерентность и однородность последнего уровня регуляризатор не влияет, поэтому графики для этих критериев не приводятся.*



Однородность промежуточного уровня



Плотность Ψ^1

Рис. 6: *Исследование влияния коэффициента p разреживания графа на качество модели. По оси абсцисс — номер итерации алгоритма обучения, по оси ординат — метрика качества. На правдоподобие, когерентность и однородность последнего уровня регуляризатор не влияет, поэтому графики для этих критериев не приводятся.*

Поскольку разреживание графа не портит другие показатели и при этом делает иерархию более похожей на дерево, сделаем вывод, что оно положительно влияет на модель.

Отметим, что без применения разреживания граф также становится похожим на дерево в том смысле, что большинство значений в матрицах Ψ^l малы относительно максимальных значений в столбцах и строках, но не равны нулю.

4.4 Выводы

Подытожим основные выводы этого раздела:

1. Благодаря совмещению в единой вероятностной модели лучших черт уже существующих подходов удается построить тематическую иерархию, удовлетворяющую всем требованиям, включая желаемые, однако без одного необходимого — требования масштабируемости.

2. С помощью эвристического приема разреживания распределений над терминами удается сделать темы более интерпретируемыми и понятными для человека, однако при этом ухудшаются правдоподобие модели и качество кластеризации.

3. В результате применения эвристического приема разреживания графа он практически превращается в дерево, при этом другие показатели качества модели не ухудшаются.

4. Разреживание модели лучше начинать после того, как модель достаточно хорошо сошлась. Можно применять стандартный параметр разреживания $p = 2$.

5 Экспериментальное сравнение двух предложенных подходов

Коллекции. Сравнение двух подходов производилось на двух коллекциях. Первая — коллекция школьных конспектов. Вторая коллекция — набор текстовых записей видеолекций, скачанных с сайта Постнаука[2]. В коллекции 1728 документов, 38467 слов. Коллекция лемматизирована и приведена к виду мешка слов. Кроме того, из словаря исключены предлоги, в отличие от коллекции школьных конспектов.

На обеих коллекциях строится иерархия с 2 уровнями, 10 и 30 тем. Для иерархии, которая строится вторым методом, также добавляется корневая тема.

Критерии качества. Сравнение производится по четырем критериям: усредненная по всем темам когерентность, плотность матрицы перехода между уровнями:

$$G_2 = \frac{1}{|T||S|} \sum_t \sum_s [\psi_{ts} > \epsilon],$$

[2]http://postnauka.ru

плотность распределений терминов в темах:

$$G_3 = \sum_{l=1}^{L} \sum_{s^l \in S^l} \sum_{w \in W} [\phi^l_{ws^l} > 0]$$

и время обучения модели. Порог ϵ вводится по той причине, что хотя в процессе обучения ни один элемент матрицы Ψ не обнулится, если не применять разреживание Ψ, многие значения в ней будут близки к нулю, и такие ребра не будут показываться в навигаторе. В экспериментах принимается $\epsilon = 0.05$.

Описание и результаты эксперимента. Для построения модели с регуляризатором связи уровней был выбран регуляризатор матрицы Φ, так как он лучше показал себя в экспериментах. На каждой коллекции построено две модели: с регуляризатором Φ и единая генеративная модель. Для обучения каждой модели сначала выполняется 25 итераций без разреживания, затем 25 итераций с разреживанием распределений терминов в темах. Для ARTM коэффициент регуляризации выбирается таким, при котором темы получаются наиболее интерпретируемы. Для единой модели параметр разреживания в соответствии с результатами предыдущих экспериментов устанавливается $p = 2$. Параметр при регуляризаторе связи уровней $\lambda = 10^3$ выбран из эвристических соображений.

Эксперименты проводились на ноутбуке с ОС Windows8.1, процессором Intel Core i7, тактовой частотой 2400 МГц, 8 Гб ОЗУ. Обучение обеих моделей реализовано на языке python2.7.

Результаты сравнения на коллекции школьных конспектов:

	Ср. когерентность	Плотность Ψ	Ср. плотность Φ	Время обучения
ARTM с рег. Φ	1.166	**0.129**	0.056	**10 мин. 22 с**
Единая модель	**1.223**	0.433	**0.049**	30 мин. 34 с

Результаты сравнения на коллекции материалов Постнауки:

	Ср. когерентность	Плотность Ψ	Ср. плотность Φ	Время обучения
ARTM с рег. Φ	**1.253**	**0.146**	**0.055**	**17 мин. 14 с**
Единая модель	1.007	0.46	0.057	31 мин. 59 с

Анализ результатов. На коллекции школьных конспектов вторая модель показывает чуть лучшую интерпретируемость тем по критериям когерентности и плотности распределений. На второй коллекции ARTM показывает значительно лучшую когерентность. Причина расхождения в том, что в словаре первой коллекции присутствует много служебных слов, и вторая модель собирает их в корневой теме. В ARTM реализовать корневую тему нельзя, потому что тематическая модель с одной темой не имеет смысла. На коллекции с предварительно фильтрованным словарем ARTM строит более интерпретируемые темы.

Первая модель строит гораздо более разреженный, а значит, удобный для пользователя граф, чем вторая. Кроме того, она обучается в 2—3 раза быстрее второй.

Таким образом, ARTM с регуляризатором связи уровней строит более удобные для составления тематического навигатора модели.

6 Заключение

В работе рассмотрена задача автоматического построения иерархических тематических моделей, на основе которых удобно составлять интерактивные навигаторы по коллекции. Это свойство модели формализовано в виде требований к ней. Изучены существующие методы построения иерархий, в каждом из которых понятие иерархической модели трактуется по-своему, и ни одно из них не удовлетворяет всем введенным требованиям. Для решения задачи предложены два подхода: первый — нисходящий, строящий иерархию от крупных тем к узко специализированным, второй — настраивающий параметры всех уровней одновременно. Экспериментально показано, что первый подход больше подходит для составления иерархических тематических навигаторов.

Оба подхода позволяют строить иерархии с разреженными темами, допускающие множественное наследование тем и предоставляющие средства разреживания тематического графа. В отличие от первого подхода, единая генеративная модель позволяет добавлять в граф иерархии корневую вершину; это важно, если в словаре присутствует много служебных слов. Кроме того, эта модель более гибкая: из нее можно извлекать различные представления списков документов и терминов темы. В ARTM также можно создать отдельную тему, собирающую слова общей лексики, однако для этого придется настраивать несколько дополнительных гиперпараметров модели. В целом построение единой модели более автоматизировано: необходимо задать только количество тем на каждом уровне и траекторию разреживания; в ARTM нужно также указывать коэффициенты регуляризации. С другой стороны, в ARTM модель строится быстрее.

К сожалению, невозможно удовлетворить все возможные пожелания к модели. Такие приоритеты, предложенные в других работах, как автоматическое выделение словосочетаний, независимость настраиваемых параметров от начального приближения, автоматическое определение количества уровней, возможность достраивания модели при добавлении в коллекцию новых документов не учитываются в предложенных подходах. С другой стороны, такие идеи, как автоматическое настраивание количества тем на уровне или возможность коррекции иерархии по указанию эксперта реализуемы в первой модели благодаря широким возможностям регуляризации в ARTM.

Напоследок отметим, что несмотря на многообразие подходов к построению темати-

ческих иерархий, до сих пор остается открытым вопрос их автоматического оценивания. Для проверки качества модели приходится вводить несколько критериев качества, но они, тем не менее, не покрывают всех свойств, которыми должна обладать удобная и понятная человеку иерархия. Наиболее правильным в литературе считается способ экспертного оценивания моделей, однако это наиболее ресурсозатратный метод.

На защиту выносятся следующие результаты:

1. Предложены две нисходящие стратегии построения тематической иерархии с использованием регуляризатора связи уровней.

2. Предложена иерархическая вероятностная тематическая модель текстовой коллекции, основанная на сумме глубоких матричных разложений.

3. Предложена методология оценивания качества тематических иерархий.

Список литературы

[1] Воронцов К. В. Потапенко А. А. Фрей А. И. Апишев М. А. Дойков Н. В. Шапулин А. В. Чиркова Н. А. Многокритериальные и многомодальные вероятностные тематические модели коллекций текстовых документов // Интеллектуализация обработки информации (ИОИ-2014): Тезисы докл. — Торус Пресс, 2014. — С. 198–199.

[2] Чиркова Н. А. Иерархические вероятностные тематические модели // Труды 57-й научной конференции МФТИ с международным участием, посвященной 120-летию со дня рождения П. Л. Капицы. — 2014. — С. 18.

[3] Чиркова Н. А. Иерархические вероятностные тематические модели // Сборник тезисов XXII Международной научной конференции студентов, аспирантов и молодых учёных «Ломоносов-2015», секция «Вычислительная математика и кибернетика». — 2015. — С. 74–76.

[4] Чиркова Н. А. Айсина Р. М. Воронцов К. В. Иерархическая аддитивно регуляризованная тематическая модель научной конференции // Тезисы докладов 17-й Всероссийской конференции «Математические методы распознавания образов». — Торус Пресс, 2015. — С. 230–231.

[5] The Author-topic Model for Authors and Documents / Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, Padhraic Smyth // Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence. — UAI '04. — Arlington, Virginia, United States : AUAI Press, 2004. — С. 487–494.

[6] Automatic Taxonomy Construction from Keywords / Xueqing Liu, Yangqiu Song, Shixia Liu, Haixun Wang // Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. — KDD '12. — New York, NY, USA : ACM, 2012. — С. 1433–1441.

[7] Blei David M., Ng Andrew Y., Jordan Michael I. Latent Dirichlet Allocation // J. Mach. Learn. Res. — 2003. — Март. — Т. 3. — С. 993–1022.

[8] Chemudugunta Chaitanya, Smyth Padhraic, Steyvers Mark. Modeling General and Specific Aspects of Documents with a Probabilistic Topic Model // Advances in Neural Information Processing Systems 19 / Под ред. B. Schölkopf, J. Platt, T. Hoffman. — Cambridge, MA : MIT Press, 2006. — С. 241–248.

[9] Constructing topical hierarchies in heterogeneous information networks / Chi Wang, Jialu Liu, Nihit Desai и др. // Knowledge and Information Systems. — 2014. — Т. 44, № 3. — С. 529–558.

[10] Evaluating hierarchical organisation structures for exploring digital libraries / Mark M. Hall, Samuel Fernando, Paul D. Clough и др. // Inf. Retr. — 2014. — Т. 17, № 4. — С. 351–379.

[11] Hierarchical topic models and the nested Chinese restaurant process / David M. Blei, Thomas Griffiths, Michael Jordan, Joshua Tenenbaum // NIPS. — 2003.

[12] Hofmann Thomas. Probabilistic Latent Semantic Indexing // Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. — SIGIR '99. — New York, NY, USA : ACM, 1999. — С. 50–57.

[13] Incremental Hierarchical Clustering of Text Documents / Nachiketa Sahoo, Jamie Callan, Ramayya Krishnan и др. // Proceedings of the 15th ACM International Conference on Information and Knowledge Management. — CIKM '06. — New York, NY, USA : ACM, 2006. — С. 357–366.

[14] Li Wei, McCallum Andrew. Pachinko allocation: DAG-structured mixture models of topic correlations // ICML. — 2006.

[15] Mimno David, Li Wei, McCallum Andrew. Mixtures of Hierarchical Topics with Pachinko Allocation // ICML. — 2007.

[16] Newman David, Bonilla Edwin V., Buntine Wray L. Improving Topic Coherence with Regularized Topic Models // Advances in Neural Information Processing Systems 24:

25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain. — 2011. — С. 496–504.

[17] Non-Bayesian Additive Regularization for Multimodal Topic Modeling of Large Collections / Konstantin Vorontsov, Oleksandr Frei, Murat Apishev и др. // Proceedings of the 2015 Workshop on Topic Models: Post-Processing and Applications. — TM '15. — New York, NY, USA : ACM, 2015. — С. 29–37.

[18] Optimizing Semantic Coherence in Topic Models / David Mimno, Hanna M. Wallach, Edmund Talley и др. // Proceedings of the Conference on Empirical Methods in Natural Language Processing. — EMNLP '11. — Stroudsburg, PA, USA : Association for Computational Linguistics, 2011. — С. 262–272.

[19] A Phrase Mining Framework for Recursive Construction of a Topical Hierarchy / Chi Wang, Marina Danilevsky, Nihit Desai и др. // Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. — KDD '13. — New York, NY, USA : ACM, 2013. — С. 437–445.

[20] Pujara Jay, Skomoroch Peter. Large-Scale Hierarchical Topic Models // NIPS Workshop on Big Learning. — 2012.

[21] Rosenberg Andrew, Hirschberg Julia. V-Measure: A Conditional Entropy-Based External Cluster Evaluation Measure // Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL). — 2007. — С. 410–420.

[22] Scalable and Robust Construction of Topical Hierarchies / Chi Wang, Xueqing Liu, Yanglei Song, Jiawei Han // CoRR. — 2014. — Т. abs/1403.3460.

[23] Than Khoat, Ho Tu Bao. Fully Sparse Topic Models // Proceedings of the 2012 European Conference on Machine Learning and Knowledge Discovery in Databases - Volume Part I. — ECML PKDD'12. — Berlin, Heidelberg : Springer-Verlag, 2012. — С. 490–505.

[24] Towards Interactive Construction of Topical Hierarchy: A Recursive Tensor Decomposition Approach / Chi Wang, Xueqing Liu, Yanglei Song, Jiawei Han // Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. — KDD '15. — New York, NY, USA : ACM, 2015. — С. 1225–1234.

[25] Vorontsov Konstantin, Potapenko Anna. Analysis of Images, Social Networks and Texts: Third International Conference, AIST 2014, Yekaterinburg, Russia, April 10-12, 2014, Revised Selected

Papers / Под ред. I. Dmitry Ignatov, Yu. Mikhail Khachay, Alexander Panchenko и др. — Cham : Springer International Publishing, 2014. — С. 29–46. — ISBN: 978-3-319-12580-0.

[26] Xiao Han, Stibor Thomas. Efficient Collapsed Gibbs Sampling For Latent Dirichlet Allocation // Asian Conference on Machine Learning (ACML). — Т. 13 из JMLR W&CP. — Japan, 2010. — (AR: 31

[27] Yee Whye Teh Michael I. Jordan Matthew J. Beal David M. Blei. Hierarchical Dirichlet Processes // Journal of the American Statistical Association. — 2006. — Т. 101, № 476. — С. 1566–1581.

[28] Zavitsanos Elias, Paliouras Georgios, Vouros George A. Non-Parametric Estimation of Topic Hierarchies from Texts with Hierarchical Dirichlet Processes // J. Mach. Learn. Res. — 2011. — Нояб. — Т. 12. — С. 2749–2775.