ПРИЛОЖЕНИЕ «А»

к отчету о научно-исследовательской работе

«Мастерская знаний: создание систем на основе гибридного ИИ для управления знаниями в научной деятельности» (промежуточный этап 1)

Обзор математической теории и методов аддитивной регуляризации тематических моделей ARTM и новые тематические модели локальных контекстов

Обзор подготовлен в рамках научно-исследовательских работ в области искусственного интеллекта по теме «Мастерская знаний: создание систем на основе гибридного ИИ для управления знаниями в научной деятельности» исследовательского центра МГУ в сфере искусственного интеллекта, финансируемого в соответствии с решением о порядке предоставления субсидии №25-66879-02035-Р от 6 февраля 2025 г. Министерства экономического развития Российской Федерации.

Раздел 1 отражает состояние исследований по вероятностному тематическому моделированию в рамках подхода аддитивной регуляризации (additive regularization for topic modeling, ARTM). Акцент сделан на достижение полной интерпретируемости всех тем в тематической модели. Дефицит интерпретируемости снижает прикладную ценность тематического моделирования, в частности, для разведочного анализа больших коллекций научной периодики. До сих пор данная проблема не имела окончательного решения ни в одном из существующих подходов (байесовском, нейросетевом, регуляризационном), несмотря на значимые продвижения в отдельных направлениях. Тем не менее, именно в рамках ARTM складываются предпосылки для решения проблемы дефицита интерпретируемости.

В разделе 2 предлагается новое семейство регуляризованных тематических моделей, основанное на анализе локальных контекстов слов. Учёт локальных контекстов оказывает регуляризирующее воздействие на модель, которое способствует повышению интерпретируемости. При этом в модель вводится возможность управления балансировкой тем в процессе её обучения.

1 Обзор математической теории и методов

1.1 Задача вероятностного тематического моделирования

Тематическое моделирование как технология *обработки естественного языка* (natural language processing, NLP) активно развивается с конца 90-х годов [?, ?].

Тематическая модель коллекции текстовых документов определяет, к каким темам относится каждый документ, и какие слова образуют каждую тему.

Тематическое моделирование принято относить к *машинному обучению без учителя* (unsupervised learning) и к *мягкой кластеризации текстов* (soft clustering of text), поскольку модель находит кластеры-темы в исходных данных автоматически, без использования каких-либо подсказок — разметок, тезаурусов или баз знаний.

Вероятностная тематическая модель (probabilistic topic model, PTM) определяет по текстовой коллекции два семейства вероятностных распределений — вероятности тем для каждого документа и вероятности слов для каждой темы.

Распределение вероятностей тем используется в качестве *векторного представления* (embedding) текста при решении задач информационного поиска, классификации, сегментации, суммаризации документов или их фрагментов.

Распределение вероятностей слов позволяет выделить в теме наиболее частотные слова. Если они оказываются связанными друг с другом по смыслу, то их называют *семантическим ядром* темы, а тему называют *интерпретируемой*. Для такой темы эксперт без труда подберёт название и составит текстовое описание [?, ?].

Интерпретируемость важна для большинства применений тематического моделирования в разведочном информационном поиске (exploratory search) [?, ?, ?, ?], в том числе в поиске научно-технической информации, в цифровых гуманитарных исследованиях (digital humanities), текстовой аналитике в области филологии, культурологии, психологии, социологии, политологии, истории, журналистики, маркетинга [?, ?, ?, ?]. В типичных случаях тематическая модель помогает понять, «о чём все эти тексты», не читая их, или отфильтровать ту часть коллекции, которая представляет интерес для более детального анализа. При этом исследователи могут не знать заранее, какие именно темы или их аспекты окажутся значимыми. Тематическое моделирование является инструментом для извлечения новых знаний и нетривиальных выводов из больших текстовых данных.

Байесовское обучение надолго стало основной математической техникой в тематическом моделировании после работ [?, ?]. В байесовском подходе оцениваются не сами параметры модели, а их апостериорные распределения. Такое усложнение на практике оказывается избыточным. *Аддитивная регуляризация тематических моделей* (ARTM), предложенная в [?], основана на более простой, но не менее выразительной оптимизационной технике, позволяющей комбинировать модели для получения требуемых свойств.

Цель данного обзора — представить ретроспективу развития и современное состояние теории и методов ARTM, наметить пути решения открытых проблем тематического моделирования.

1.2 Тематические модели PLSA и LDA

Постановка задачи вероятностного тематического моделирования была впервые предложена Томасом Хофманном для модели *вероятностного латентного семантического анализа* (probabilistic latent semantic analysis, PLSA) [?, ?].

Пусть заданы три конечных множества: D — коллекция текстовых документов, W — словарь термов, T — множество тем. Каждый документ $d \in D$ представляет собой последовательность n_d термов $(w_1, \ldots, w_{n_d})$ из словаря W. Обычно в роли термов выступают слова, но также это могут быть словосочетания или части слов, в зависимости от используемых методов предварительной обработки текста.

Тематическая модель PLSA, как и большинство последующих, основана на *гипотезе «мешка слов»*. Документ представляется мультимножеством термов $d \subset W$, в котором каждый терм $w \in d$ встречается n_{dw} раз, при этом информация о порядке термов в документе игнорируется.

Модель PLSA оценивает условные вероятности $p(w \,|\, d)$ появления термов w в документах d через вероятности термов в темах $\varphi_{wt} = p(w \,|\, t)$ и вероятности тем в документах $\theta_{td} = p(t \,|\, d)$. Модель опирается на *гипотезу условной независимости* — предположение, что появление термов темы t в документе d зависит от темы, но не зависит от документа, $p(w \,|\, d, t) = p(w \,|\, t)$. Согласно формуле полной вероятности, из этих предположений следует, что

$$p(w \,|\, d) = \sum_{t \in T} p(w \,|\, d, t)\, p(t \,|\, d) = \sum_{t \in T} \varphi_{wt} \theta_{td}. \tag{1}$$

Параметры тематической модели PLSA — матрицы $\Phi = (\varphi_{wt})_{W \times T}$ и $\Theta = (\theta_{td})_{T \times D}$ оцениваются по коллекции D путём максимизации правдоподобия:

$$\sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \varphi_{wt} \theta_{td} \;\to\; \max_{\Phi, \Theta}, \tag{2}$$

при ограничениях неотрицательности и нормировки на столбцы матриц Φ и Θ, которые являются дискретными вероятностными распределениями:

$$\sum_{w \in W} \varphi_{wt} = 1, \quad \varphi_{wt} \geqslant 0; \qquad \sum_{t \in T} \theta_{td} = 1, \quad \theta_{td} \geqslant 0. \tag{3}$$

Оптимизационная задача (2)–(3) представляет собой задачу низкорангового стохастического матричного разложения. Она относится к классу некорректно поставленных задач, поскольку имеет в общем случае бесконечное множество решений. Чтобы доопределить постановку задачи, необходимо вводить регуляризацию — дополнительные ограничения на матрицы Φ и Θ.

Дэвид Блэй, Эндрю Ын и Майкл Джордан в [?, ?] ввели модель *латентного размещения Дирихле* (latent Dirichlet allocation, LDA), основанную на предположении, что столбцы матриц Φ и Θ порождаются априорными распределениями Дирихле размерности $|W|$ и $|T|$ соответственно. Модель LDA до сих пор остаётся наиболее цитируемой в тематическом моделировании. Многие последующие модели были развитием и специализацией LDA для конкретных ситуаций в рамках математической теории байесовского обучения [?, ?, ?, ?].

1.3 От байесовской регуляризации к аддитивной

Обозначим для общности через X исходные данные, $\Omega = (\Phi, \Theta)$ — параметры модели, γ — гиперпараметры априорного распределения.

Методы байесовского вывода позволяют оценить апостериорное распределение параметров $p(\Omega \mid X, \gamma)$ по заданному априорному распределению $p(\Omega \mid \gamma)$:

$$p(\Omega \mid X, \gamma) = \frac{p(X \mid \Omega)\, p(\Omega \mid \gamma)}{\int p(X \mid \Omega)\, p(\Omega \mid \gamma)\, d\Omega}.$$

Апостериорное распределение полезно для вероятностного анализа модели: проверки статистических гипотез, получения доверительных интервалов, анализа устойчивости и т. д. Однако в практике тематического моделирования оно используется исключительно ради получения точечной оценки максимума правдоподобия:

$$\Omega = \arg\max_{\Omega} p(\Omega \mid X, \gamma).$$

Метод *максимума апостериорной вероятности* (maximum a posteriori probability, MAP) позволяет получить ту же точечную оценку непосредственно, минуя байесовский вывод, связанный с трудоёмким интегрированием:

$$\Omega = \arg\max_{\Omega} \big(\ln p(X \mid \Omega) + \ln p(\Omega \mid \gamma)\big).$$

Здесь первое слагаемое совпадает с (2), второе играет роль дополнительного оптимизационного критерия — *регуляризатора*. В общем случае регуляризатор не обязан быть логарифмом априорного распределения или иметь какой-либо вероятностный смысл. Более того, для агрегирования различных требований к модели можно брать взвешенную сумму нескольких критериев регуляризации:

$$\Omega = \arg\max_{\Omega} \big(\ln p(X \mid \Omega) + \sum_i \tau_i R_i(\Omega)\big),$$

где τ_i — коэффициенты регуляризации, отвечающие за балансировку критериев. Их подбирают, как правило, экспериментальным путём.

Различия между байесовским подходом и оптимизационным представляются на первый взгляд принципиальными. Однако получаемые в результате итерационные алгоритмы оказываются очень похожими. Впервые этот факт был замечен, по всей видимости, в [?] при сравнении алгоритмов обучения моделей PLSA и LDA.

Сравнение моделей PLSA, LDA с их робастными аналогами (special words with background, SWB), в которых выделяются фоновые и специфичные слова, показало, что преимущество LDA над PLSA несколько переоценено [?, ?, ?]. Правдоподобие модели LDA действительно выше, но это достигается благодаря смещению оценок вероятности для редких специфичных слов, которые как раз наименее важны в темах и не влияют на интерпретируемость модели. Эти исследования показали необходимость более сильной проблемно-ориентированной регуляризации.

Аддитивная регуляризация тематических моделей (additive regularization for topic modeling, ARTM) была предложена Константином Воронцовым в [?]. Первоначально рассматривалась регуляризация модели PLSA (2)–(3):

$$\sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{t\in T} \varphi_{wt}\theta_{td} + R(\Phi,\Theta) \;\to\; \max_{\Phi,\Theta}, \quad R(\Phi,\Theta) = \sum_i \tau_i R_i(\Phi,\Theta). \tag{4}$$

Позже аддитивная регуляризация была распространена и на более сложные виды тематических моделей, о которых пойдёт речь ниже. В общем случае регуляризаторы могут зависеть от данных X, тогда корректнее говорить о многокритериальной оптимизации модели путём *скаляризации критериев*.

ARTM — это не отдельная модель или метод, а математическая теория, общий подход к построению и агрегированию тематических моделей, простая и гибкая альтернатива байесовскому обучению. Формулы итерационного процесса в ARTM выводятся для любых тематических моделей единообразно, как следствия из общей теоремы о максимизации гладкой функции на единичных симплексах [**?**].

Лемма 1. *[?]. Пусть функция $f(\Omega)$ непрерывно дифференцируема по набору векторов $\Omega = (\omega_j)_{j\in J}$, $\omega_j = (\omega_{ij})_{i\in I_j}$. Если ω_j — вектор локального экстремума задачи математического программирования*

$$f(\Omega) \to \max_{\Omega}, \quad \sum_{i\in I_j} \omega_{ij} = 1, \quad \omega_{ij} \geqslant 0, \quad i \in I_j,\ j \in J$$

и если $\omega_{ij}\frac{\partial f}{\partial \omega_{ij}} > 0$ при некотором i, то ω_j удовлетворяет уравнениям

$$\omega_{ij} = \underset{i\in I_j}{\operatorname{norm}}\left(\omega_{ij}\frac{\partial f}{\partial \omega_{ij}}\right). \tag{5}$$

Операция norm преобразует произвольный вектор $(x_i)_{i\in I}$ в неотрицательный нормированный вектор $(p_i)_{i\in I}$ дискретного вероятностного распределения (если $x_i \leqslant 0$ для всех $i \in I$, то результат norm определяется как нулевой вектор):

$$p_i = \underset{i\in I}{\operatorname{norm}}(x_i) = \frac{\max(0, x_i)}{\sum\limits_{j\in I}\max(0, x_j)}, \text{ для всех } i \in I.$$

Теорема 1. *[?]. Пусть функция $R(\Phi, \Theta)$ непрерывно дифференцируема. Тогда локальным экстремумом задачи (4) с ограничениями (3) являются матрицы (Φ, Θ), ненулевые столбцы которых удовлетворяют следующей системе уравнений со вспомогательными переменными $p_{tdw} = p(t \mid d, w)$, для всех $d \in D$, $w \in W$, $t \in T$:*

$$p_{tdw} = \underset{t\in T}{\operatorname{norm}}\big(\varphi_{wt}\theta_{td}\big); \tag{6}$$

$$\varphi_{wt} = \underset{w\in W}{\operatorname{norm}}\left(n_{wt} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}}\right); \qquad n_{wt} = \sum_{d\in D} n_{dw}p_{tdw}; \tag{7}$$

$$\theta_{td} = \underset{t\in T}{\operatorname{norm}}\left(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right); \qquad n_{td} = \sum_{w\in d} n_{dw}p_{tdw}. \tag{8}$$

Систему уравнений (6)–(8) удобно решать методом простых итераций, который принято называть EM-алгоритмом (expectation–maximization). Сначала выбираются начальные приближения $\varphi_{wt}, \theta_{td}$, затем в цикле до сходимости чередуются *E-шаг* (6) и *M-шаг* (7)–(8).

Интерпретации вспомогательных переменных как частотных оценок:

n_{wt} — сколько раз терм $w \in W$ относится к теме t во всей коллекции;

n_{td} — сколько раз термы документа $d \in D$ относятся к теме t.

Появление нулевого столбца в матрице Φ интерпретируется как исключение из модели соответствующей (аномальной) темы. Нулевой столбец в матрице Θ означает отказ модели от тематизации данного (аномального) документа. Нулевые столбцы могут появляться только в результате воздействия регуляризатора. Возможность влиять на структуру модели является полезной особенностью ARTM.

В последующих работах [?, ?, ?] байесовские тематические модели были массово переформулированы в виде регуляризаторов $R(\Phi, \Theta) = \ln p(\Phi, \Theta \,|\, \gamma)$: разреживание и сглаживание (аналог модели LDA), частичное обучение, декоррелирование тем для повышения их различности, классификация документов и другие [?, ?, ?].

В частности, было показано, что модель LDA соответствует требованию, чтобы столбцы матрицы Φ были похожи на заданный вектор (сглаживание вероятностей) или, наоборот, не похожи (разреживание вероятностей). Аналогично вводятся регуляризаторы сглаживания и разреживания столбцов матрицы Θ. ARTM позволяет определять эти регуляризаторы без введения априорных распределений Дирихле — через критерий минимума KL-дивергенции или максимума правдоподобия. Приятным бонусом оказывается снятие ограничений на величину разреживания, продиктованных строгой положительностью векторов в распределении Дирихле [?].

Модель LDA играет особую роль в байесовском тематическом моделировании благодаря сопряжённости распределения Дирихле и мультиномиального распределения, что сильно упрощает байесовский вывод. Модель LDA не имеет убедительных лингвистических обоснований, но благодаря математическому удобству она стала общепринятой теоретической базой всего тематического моделирования. В ARTM нет оснований предпочитать распределение Дирихле другим регуляризаторам, что открывает больше возможностей для проблемно-ориентированной регуляризации.

Подход ARTM позволяет суммировать регуляризаторы от разных моделей, получая агрегированные модели с требуемым набором свойств. Все они обучаются одним и тем же итерационным алгоритмом, который реализуется и отлаживается один раз, а регуляризаторы добавляются в него как отдельные вычислительные модули. Байесовский подход не поддерживает такого рода гибкости и модульности, поскольку для каждого априорного распределения байесовский вывод уникален и приводит к новому алгоритму, который приходится заново реализовывать и отлаживать.

Илья Ирхин в теоретической работе [?] доказал сходимость итерационного алгоритма обучения для широкого класса аддитивно регуляризованных тематических моделей при весьма общих и легко проверяемых допущениях о виде регуляризатора. Там же было показано, что сходимость алгоритма к стационарной точке оптимизационной задачи можно улучшить без дополнительных затрат времени и памяти. Для этого необходимо подставить в правые части формул M-шага вместо φ_{wt} и θ_{td} несмещённые (при $R = 0$) частотные оценки этих условных вероятностей, $\varphi^*_{wt} = \operatorname{norm}_w(n_{wt})$ и $\theta^*_{td} = \operatorname{norm}_t(n_{td})$ соответственно:

$$\varphi_{wt} = \underset{w \in W}{\operatorname{norm}}\left(n_{wt} + \varphi^*_{wt}\frac{\partial R(\Phi^*, \Theta^*)}{\partial \varphi_{wt}}\right); \tag{9}$$

$$\theta_{td} = \underset{t \in T}{\operatorname{norm}}\left(n_{td} + \theta^*_{td}\frac{\partial R(\Phi^*, \Theta^*)}{\partial \theta_{td}}\right). \tag{10}$$

Эксперименты в [?] показали, что данная модификация повышает значение регуляризованного правдоподобия, причём итерационный процесс гораздо быстрее (как

правило, уже на второй итерации) входит в режим монотонного увеличения регуляризованного правдоподобия. Чем больше коэффициент регуляризации, тем сильнее проявляется эффект улучшения сходимости.

Анна Потапенко в серии экспериментов [?, ?, ?] показала, что агрегирование шести регуляризаторов сглаживания, разреживания, декоррелирования и отбора тем улучшает одновременно несколько метрик качества (когерентность, разреженность, чистоту и контрастность тем) ценой незначительного ухудшения правдоподобия. При агрегировании регуляризаторы могут включаться как параллельно, так и последовательно. Управляемое изменение коэффициентов регуляризации τ_i в ходе итерационного процесса называется *стратегией регуляризации*. Некоторые эмпирические правила по выстраиванию стратегии регуляризации были выработаны в [?]: декоррелирование и сглаживание лучше включать с первой же итерации, разреживание — после образования тенденции к сходимости (обычно через 5–20 итераций, в зависимости от размера коллекции). Нельзя разреживать матрицы Φ и Θ сразу, так как небольшая доля параметров, случайно оказавшихся близкими к нулю, могут так и остаться нулевыми. Эти рекомендации использовались во многих последующих работах для улучшения качества модели [?, ?, ?].

Не существует теоретических или практических препятствий к тому, чтобы переписать в терминах ARTM всю теорию вероятностного тематического моделирования.

Возможно, переход от байесовского вывода к более простой градиентной оптимизации поспособствует дальнейшему симбиозу вероятностных тематических моделей с современными нейросетевыми моделями языка. До сих пор попытки их совмещения наталкивались на значительные технические трудности [?].

1.4 От LDA к проблемно-ориентированной регуляризации

Преимущества ARTM проявляются во многих приложениях, где тематическая модель используется как вспомогательный инструмент для решения смежных задач текстовой аналитики.

Тематические векторы $p(t \mid \cdot)$ часто используются в качестве признаковых описаний текстовых объектов в задачах информационного поиска и предсказательного моделирования — классификации, регрессии, ранжирования. Потребность в предсказании числовой величины по текстовому документу возникает во многих приложениях [?]: предсказание рейтинга товара, фильма или книги по отзыву; предсказание числа кликов по рекламному объявлению; предсказание зарплаты по описанию вакансии; предсказание полезности (числа лайков) отзыва на отель, ресторан, сервис; предсказание изменения курса актива по экономическим новостям и т. д.

Евгений Соколов и Лев Боголюбский в [?] предложили использовать критерий наименьших квадратов, обычный для регрессионных моделей, в качестве регуляризатора для тематической модели ARTM. В итерационном процессе две модели поочерёдно приспосабливаются друг к другу. Тематическая модель формирует векторы вероятностей тем $p(t \mid d)$, которые используются в качестве признаков документов d в линейной регрессии. Регрессионная модель воздействует на темы через регуляризатор таким образом, чтобы тематические признаки оказывались как можно более информативными.

Сергей Николенко в [?] объединил тематическую модель текстовых документов с моделью сингулярного матричного разложения (singular value decomposition, SVD)

для анализа пользовательского поведения в системе рекомендаций текстового контента. В работе отмечается, что две модели, применяемые к взаимосвязанным частям данных, намного проще агрегируются в рамках ARTM, чем в байесовском подходе.

Мурат Апишев в [?, ?] применил ARTM для выявления тематики национальностей и межнациональных отношений в постах социальных сетей. Для поиска этнорелевантных тем социологами был составлен словарь из нескольких сотен этнонимов. Были предложены два способа словарного воздействия на модель в рамках ARTM. Первый способ — применять регуляризатор сглаживания по словам словаря к заранее выделенному целевому подмножеству тем. Второй способ — использовать словарь как отдельную модальность, придавая словам словаря большой вес. В обоих случаях тематическая модель автоматически определяла, как именно этничности распределяются по темам. По сравнению с предыдущими исследованиями [?, ?] подход ARTM помогал социологам находить полиэтничные темы и выявлять особенности межэтнических отношений.

Мария Милкова в [?] использовала 22 регуляризатора сглаживания для выделения 22 тем, соответствующих отраслям экономики, для анализа программы импортозамещения на основе патентных данных.

В [?] строилась тематическая модель по 13.6M репозиториев программного обеспечения на GitHub с целью понимания и категоризации тематики открытых проектов. Были построены 256 тем, которые удалось распределить вручную по 6 крупным категориям. Для повышения интерпретируемости тем использовались регуляризаторы разреживания.

Многие приложения тематического моделирования связаны с обработкой динамически пополняемых текстовых коллекций, таких как социальные медиа, новостные потоки, научные архивы, патентные базы.

Николай Герасименко и др. в [?, ?, ?] разработали инкрементную тематическую модель с аддитивной регуляризацией iARTM для автоматического выделения новых научных трендов. Тематическая модель последовательно дообучалась на новых пакетах документов, которые проверялись на появление новой терминологии и новых тем. Тренд определяется как семантически однородная тема с устойчивым лексическим ядром, объём которой резко возрастает в определённый момент времени. Тренд характеризуется стартовой публикацией и ключевыми терминами, среди которых, как правило, выделяется название тренда.

В серии публикаций [?, ?, ?] были предложены способы построения динамических тематических моделей ARTM для мониторинга социальных медиа с целью информационной поддержки регионального и муниципального развития. Выходные данные тематического моделирования объектов и процессов социальных медиа использовались для дальнейшего статистического анализа и когнитивной визуализации результатов динамического тематического моделирования.

В [?] тематическая модель ARTM строилась для изучения нарративов информационной войны между Россией и Украиной по данным Твиттера до февраля 2022 года. Для увеличения числа интерпретируемых тем использовались три регуляризатора. Первый — декоррелирование для увеличения различности тем. Второй — разреживание тематических векторов документов для исключения нерелевантных тем, исходя из предположения, что каждый документ относится к небольшому числу тем. Третий — сглаживание тематических распределений соседних временных

интервалов (продолжительностью 1 неделя) для обеспечения непрерывности тематики во времени, обнаружения новых тем и их дальнейшего отслеживания.

1.5 От теории ARTM к реализации BigARTM

Александр Фрей и Мурат Апишев на основе теории ARTM разработали библиотеку `BigARTM` с открытым кодом [?, ?]. В ней реализован многопоточный пакетный онлайновый алгоритм для эффективной обработки больших текстовых коллекций.

При *пакетной* обработке исходные данные разбиваются на пакеты, которые по очереди загружаются в оперативную память, обрабатываются, и затем выгружаются, освобождая место для следующего пакета. В некоторых случаях исходные данные настолько большие (миллиарды термов), что итерационный процесс успевает сойтись, обработав все пакеты только по одному разу. В таких случаях мы говорим про *онлайновый* алгоритм. Он особенно удобен, когда новые документы продолжают поступать в коллекцию динамически. Статические и относительно небольшие текстовые коллекции (менее миллиона термов), как правило, требуют нескольких повторных проходов всей коллекции. Такие алгоритмы называются *оффлайновыми*. `BigARTM` предоставляет оба варианта реализации, онлайновый и оффлайновый. Число проходов, размер пакета, частота обновления матрицы Φ являются управляющими параметрами алгоритма обучения тематической модели в `BigARTM`.

Библиотека является кроссплатформенной, имеет встроенные и расширяемые наборы регуляризаторов и метрик качества, поддерживает формат текстовых данных Vowpal Wabbit. Ядро библиотеки написано на языке `C++`, программный интерфейс — на языке Python. Полная документация находится на сайте `http://bigartm.org`.

Наиболее эффективный асинхронный алгоритм распараллеливания вычислений для `BigARTM` описан в [?]. Там решена проблема равномерной загрузки процессоров и показано, что скорость вычислений практически не деградирует при увеличении числа ядер процессора. Эти результаты были усилены М. Апишевым в [?] и показано, что алгоритм `BigARTM` во много раз превосходит конкурирующие библиотеки с открытым кодом Gensim и Vowpal Wabbit по скорости вычислений. При использовании 8 процессоров асинхронный алгоритм показал преимущество в 28 раз.

Ещё один способ ускорения вычислений предложен в [?]. Функция логарифма в постановке задачи (2) заменяется произвольной гладкой возрастающей функцией. Алгоритм обучения тематической модели модифицируется не сильно, но в одном частном случае из него уходит дорогостоящая операция нормировки тематического вектора для каждого слова в каждом документе. Это приводит к быстрым, но неточным вычислениям, которые можно использовать на начальных итерациях, в целом ускорив итерационный процесс в полтора раза без потери итогового качества модели.

В обстоятельном обзоре М. Апишева [?] сравниваются эффективные алгоритмы обучения вероятностных тематических моделей LDA и ARTM для многопроцессорных систем; предлагается систематизация технических приёмов для организации параллельных вычислений, распределённого хранения данных, потоковой обработки, уменьшения потребления оперативной памяти, повышения отказоустойчивости.

1.6 От выбора числа тем к тематическим иерархиям

Вопрос о выборе числа тем возникает в каждой практической задаче тематического моделирования. Внутри любой темы могут обнаружиться вариации или аспекты, но достаточно ли между ними различий, чтобы разделить тему на подтемы или это частные проявления одной общей темы? Существуют ли критерии для различения этих ситуаций и определения оптимального числа тем?

Александр Плавин в [?] исследовал возможность применения энтропийного регуляризатора разреживания, ранее предложенного в [?], для удаления незначимых тем. В байесовском подходе для определения числа тем используют непараметрическую модель иерархического процесса Дирихле (hierarchical Dirichlet process, HDP) [?]. В обеих моделях имеется параметр, от которого число тем зависит существенно: коэффициент регуляризации τ в ARTM и коэффициент концентрации γ в HDP.

Эксперименты в [?] проводились на полусинтетических данных, генерируемых смесью двух распределений $p(w\,|\,d)$ — реальной коллекции, для которой число тем неизвестно, и синтетической коллекции, сгенерированной тематической моделью с заданным числом тем, предварительно построенной по той же коллекции. Оказалось, что алгоритмы HDP и ARTM неплохо определяют истинное число тем на синтетических и полусинтетических данных. При этом ARTM определяет его точнее и устойчивее. Однако чем ближе полусинтетические данные к реальным, тем менее чётко различим диапазон значений гиперпараметров τ или γ, на котором восстанавливается правильное число тем. На реальных данных он неразличим вовсе, так что оба подхода не в состоянии определить оптимальное число тем.

По скорости вычислений `BigARTM` с регуляризатором отбора тем оказался в 100 раз быстрее свободно доступной реализации HDP на языке `C++`.

В экспериментах [?] также выяснилось, что регуляризатор отбора тем имеет полезный сопутствующий эффект, удаляя из модели дублирующие, расщеплённые и линейно зависимые темы. Теоретическое обоснование этого эффекта пока остаётся открытой проблемой.

Оптимальное число тем трудно определить из-за того, что любая тема может дробиться на более мелкие подтемы, и предел такого дробления упирается лишь в размер коллекции. Сравнительное исследование [?] подтверждает, что число тем зависит скорее от модели и метода оценивания, чем является объективной характеристикой текстовой коллекции.

Вместо оптимизации числа тем имеет смысл строить тематические иерархии, в которых темы разделяются на подтемы по необходимости, а уровень детализации выбирается исходя из оценок качества решения целевой прикладной задачи, например поиска или классификации текстов. Падение качества может происходить как при недостаточном числе тем — из-за деградации модели, так и при избыточном числе тем — из-за недостатка данных и переобучения модели.

Стратегии построения тематических иерархий весьма разнообразны [?]. Различаются стратегии нисходящие (разделяющие) и восходящие (объединяющие), представляющие иерархию деревом или многодольным графом, разделяющие на подтемы отдельные вершины или целиком весь уровень, основанные на кластеризации документов или термов. Нельзя назвать какую-то из стратегий предпочтительной; каждая имеет свои достоинства и недостатки.

Надежда Чиркова в [?] предложила нисходящую поуровневую стратегию в рамках ARTM. На каждом следующем уровне число тем увеличивается в несколько раз и строится обычная «плоская» тематическая модель. Для моделирования связей между уровнями в модель вводятся новые параметры $p(s \mid t)$ — условные вероятности подтем s в темах t. При построении дочернего уровня родительские темы $p(w \mid t)$ рассматриваются как псевдодокументы, которые просто добавляются к коллекции. Искомые распределения $p(s \mid t)$ считываются из столбцов матрицы Θ, соответствующих псевдодокументам. Время построения иерархии остаётся линейным по объёму коллекции. Позже этот подход был улучшен в [?], реализован в `BigARTM` и использован во многих исследованиях [?, ?, ?, ?].

1.7 От слов к модальностям

Мультимодальная тематическая модель описывает документы, содержащие метаданные наряду с основным текстом. Метаданные помогают более точно определять тематику документов, и, наоборот, тематическая модель может использоваться для выявления семантики метаданных или предсказания пропущенных метаданных.

Каждый тип метаданных образует отдельную *модальность* со своим словарём. Слова, словосочетания, n-граммы, теги, названия, аббревиатуры — это примеры текстовых модальностей. В мультиязычных тематических моделях модальностями являются языки. Примерами нетекстовых модальностей являются категории, время, ссылки, авторы, пользователи и др.

Все перечисленные выше случаи, несмотря на разнообразие приложений и интерпретаций, описываются в ARTM единым формализмом модальностей [?, ?, ?]. Каждый документ рассматривается как контейнер, содержащий термы различных модальностей, включая обычные слова.

Мультимодальная модель строится путём максимизации взвешенной суммы логарифмов правдоподобия модальностей и регуляризаторов:

$$\sum_{m \in M} \tau_m \sum_{d \in D} \sum_{w \in W^m} n_{dw} \ln \sum_{t \in T} \varphi_{wt} \theta_{td} + R(\Phi, \Theta) \;\to\; \max_{\Phi, \Theta}; \tag{11}$$

$$\sum_{w \in W_m} \varphi_{wt} = 1; \quad \varphi_{wt} \geqslant 0; \qquad \sum_{t \in T} \theta_{td} = 1; \quad \theta_{td} \geqslant 0;$$

где M — множество модальностей, W_m — словарь термов модальности m. Веса τ_m позволяют балансировать модальности с учётом их важности и частотности.

Анастасия Янина применила мультимодальную модель ARTM для тематического разведочного поиска в [?] и позже улучшила её в серии работ [?, ?, ?]. Качество тематического поиска измерялось критериями точности и полноты на основе оценок асессоров. Была составлена выборка из 100 запросов — творческих заданий разведочного поиска. Каждый запрос представлял собой длинный текст объёмом не более страницы, задающий тематику поиска. Каждое задание сначала выполнялось независимо тремя асессорами, затем системой поиска документов по близости их тематических векторов $p(t \mid d)$, затем те же асессоры выбирали релевантные документы из поисковой выдачи системы. Результаты шести раундов поиска объединялись, и этот результат предполагался достаточно полным. Данная методика позволяет, единожды собрав документы, релевантные каждому запросу, многократно оценивать качество различных тематических моделей и систем поиска.

Эксперименты проводились на двух коллекциях: 175K статей русскоязычного блога habr.ru и 760K статей англоязычного блога techcrunch.com. Алгоритм тематического поиска находил больше релевантных документов, чем асессоры, которые затрачивали 30 ± 20 минут на выполнение каждого поискового задания. Комбинирование регуляризаторов декоррелирования, разреживания и сглаживания вместе с модальностями n-грамм, авторов и категорий улучшало точность и полноту поиска от 65% до 85% после подбора числа тем $|T|$ и всех коэффициентов τ_m, τ_i.

Стратегия регуляризации строилась в три этапа [?]. Первый этап из 8 итераций включал только декоррелирование, чтобы сразу сделать темы как можно более различными. Второй этап из 8 итераций добавлял разреживание столбцов матрицы Θ, чтобы по тематическим векторам $p(t \mid d)$ чётко определились релевантные темы документов. Третий этап из 8 итераций добавлял сглаживание столбцов матрицы Φ, чтобы предотвратить вырождение тем как распределений $p(w \mid t)$. На каждом из трёх этапов коэффициент регуляризации подбирался по «грубой» сетке из четырёх значений. Качество модели контролировалось по критериям перплексии и разреженности. Оказалось, что данная стратегия, нацеленная исключительно на улучшение интерпретируемости тем, автоматически улучшала также и качество тематического поиска, хотя критерии качества поиска непосредственно не оптимизировались.

Применение иерархических тематических моделей в [?] позволило улучшить точность и полноту ещё на 5–8% при одновременном увеличении оптимального числа тем $|T|$ в несколько раз. Этот результат можно интерпретировать следующим образом: постепенное дробление тем на подтемы способствует более аккуратному разреживанию тематических векторов документов $p(t \mid d)$. Темы, которых точно нет в данном документе, получают нулевую вероятность уже на первом уровне иерархии. Поэтому мелко гранулированные подтемы нижнего уровня, которые в плоской модели могли бы быть статистически ненадёжными, наследуют нулевые вероятности от своих родительских тем. В результате поисковая выдача очищается от нерелевантных «мусорных» документов, что и означает увеличение точности поиска. Полнота поиска также немного улучшается, возможно, благодаря увеличению размерности тематических векторов. Можно сказать и так, что иерархические тематические представления документов лучше соответствует человеческой логике поиска: когда мы ищем информацию, мы сразу отсекаем то, что нам точно не подходит.

Абляционные эксперименты в [?] показали, что все регуляризаторы и модальности важны, за исключением модальности авторов. Выключение любого из критериев из постановки задачи приводит к падению точности и полноты поиска.

Даниил Фельдман в [?] предложил тематическую модель мнений. Это частный случай двухуровневой тематической иерархии, в которой верхний уровень строился обычным образом по лексике, чтобы выделить *событийные темы* в новостном потоке. Второй уровень строился по другим модальностям, чтобы разделить каждую тему не на подтемы, а на поляризованные мнения о политическом событии. Использовались модальности трёх типов: именованные объекты с позитивными и негативными тональностями, именованные объекты с их семантическими ролями, триплеты $\langle$субъект, предикат, объект$\rangle$. Абляционные эксперименты показали, что каждая из трёх модальностей важна для повышения качества выделения мнений.

Аналогичная двухуровневая иерархия использовалась в [?] для маршрутизации обращений в контактный центр. Верхний уровень строился по лексике, чтобы выделить тематику обращения. Второй уровень строился по морфологическим и син-

таксическим модальностям, чтобы выделить в каждой теме клиентские интенты, например, различая фразы «заблокируйте мне счёт» и «у меня заблокирован счёт».

Марина Суворова в [?, ?, ?] использовала модальности языков для построения мультиязычной тематической модели по текстам Википедии. Модель показала неплохую точность в задаче кросс-языкового документного поиска, когда документ-запрос задаётся на одном языке, а результатами поиска должны быть документы на другом языке. Также был построен регуляризатор двуязычного словаря, позволяющий оценивать вероятности переводов слов в зависимости от тематики документа и устранять семантическую неоднозначность слов. Например, слово «сумма» чаще переводится на английский язык как «sum» в математике и как «total» в играх и космонавтике. Включение словаря в тематическую модель лишь незначительно улучшало точность кросс-языкового поиска. Это означает, что одного лишь наличия параллельных текстов вполне достаточно для построения тематического векторного пространства, в котором каждая тема представлена на нескольких языках.

Полина Потапова в [?] строила мультиязычную тематическую модель для поиска и классификации научных текстах на 100 наиболее распространённых языках мира. Коллекция состояла из публикаций научной электронной библиотеки eLibrary и параллельных текстов Википедии. Для восполнения недостающих текстов использовалась система статистического машинного перевода Moses. Переводы каждой статьи Википедии объединялись в один документ, и языки рассматривались как модальности. Одна из тем выделялась как фоновая с помощью регуляризатора сглаживания, чтобы собрать в ней слова общей лексики. Трудность в том, что размер матрицы Φ пропорционален суммарному размеру словарей всех языков. Для сокращения словарей была применена токенизация BPE (byte-pair encoding), в результате первоначальный средний объем словаря 120K токенов удалось сократить до 11K без деградации качества модели. Размер модели сократился при этом со 128 Гб до 4.8 Гб. Качество модели оценивалось по результатам поиска переводных заимствований на тестовой выборке. При обучении модели подбиралось оптимально число тем от 10 до 1000, и наилучшие результаты были достигнуты при 125 темах. Существенный прирост качества вносило добавление модальностей рубрик УДК и ГРНТИ.

Николай Герасименко в [?] строил мультимодальную модель для документного тематического поиска схожих дел по коллекции 125K судебных актов и решений Арбитражного суда Московской области. Использовались регуляризаторы декоррелирования распределений термов в темах и разреживания распределений тем в документах. В текстах выделялись модальности юридических терминов и ссылок на нормативно-правовые акты. Были получены достаточно высокие результаты точности и полноты поиска: 91% precision@5, 22% recall@20. Низкая, на первый взгляд, полнота объясняется тем, что качество поиска оценивается лишь по верхним $k = 20$ документам поисковой выдачи, тогда как число схожих дел во многих случаях оценивается сотнями. Тематический поиск даёт около 10% прироста точности при сравнении с классической векторизацией документов TF-IDF, тематическими моделями PLSA и LDA, и простой нейросетевой моделью doc2vec. Полученные темы как распределения на термах трёх модальностей (слова, юридические термины, ссылки на нормативно-правовые акты) были оценены экспертами-юристами как хорошо интерпретируемые. При этом были отмечены прямые соответствия между некоторыми темами и типами дел, с которыми им приходилось встречаться в судебной практике.

1.8 От модальностей к транзакциям

Мультимодальные тематические модели могут применяться для анализа не только текстов, но и транзакционных данных в гораздо более широком спектре приложений. Вообще, текстовую коллекцию можно рассматривать как данные о парных взаимодействиях — транзакциях (d, w) между объектами двух конечных множеств, документами d и словами w. По этим наблюдаемым данным тематическая модель восстанавливает латентные векторные представления объектов $p(t \mid d)$, $p(t \mid w)$.

Тематический анализ банковских транзакционных данных [?] основан на их аналогии с текстами: роль документов играют клиенты банка (держатели карт, физические лица), роль слов — коды категории продавца (merchant category code, MCC). Транзакция означает факт покупки, причём не важно, какого именно товара, в каком магазине и на какую сумму. Тогда темы соответствуют паттернам потребительского поведения клиентов.

Кирилл Хрыльченко в [?] перенёс эту методику на анализ банковских транзакционных данных юридических лиц. В роли документов выступают компании, термами в документе являются контрагенты — другие компании, которые заключают с данной компанией сделки по покупке или продаже товаров или услуг. Таким образом, множество всех компаний порождает две разные модальности — покупателей и продавцов. Каждая сделка сопровождается текстом платёжного поручения, который содержит название товара или услуги. Эти названия также образуют две модальности, чтобы различать покупки и продажи. Темы соответствуют видам экономической деятельности компаний. Для интерпретации тем рассматриваются два списка — товары, которые компании покупают, и товары, которые они продают, осуществляя данный вид деятельности.

Кроме модальностей контрагентов и товаров, в данной задаче имеются также числовые данные об объёмах сделок, которые могут нести косвенную информацию о виде деятельности. Поэтому в [?] вводится понятие *числовой модальности*, и для оценивания её распределения в каждой теме используется аддитивный критерий логарифма правдоподобия.

Недостаток мультимодальной модели в том, что каждый документ (компания) представляется «мешком названий», «мешком контрагентов», и «мешком объёмов». Однако в каждой транзакции контрагент, название, время и объём сделки неразрывно связаны друг с другом и порождаются общей темой (видом деятельности). Информация об этих связях в обычной мультимодальной модели теряется.

Транзакции (взаимодействия, взаимосвязи, отношения) между тремя и более разнотипными объектами наблюдаются и в других приложениях [?]. Примеры транзакций: «покупатель b купил у продавца s товар g» в сети продаж, «пользователь u кликнул объявление b на странице s» в сети интернет-рекламы, «пользователь u написал слово w на странице блога d» в социальной сети, «клиент u вылетел из аэропорта x в аэропорт y самолётом авиакомпании a» в пассажирских авиаперевозках, «клиент u оценил фильм f в ситуативном контексте s» в рекомендательной системе. Ещё одной модальностью является время транзакции. Во всех приведённых примерах транзакция нескольких объектов не сводится к их парным взаимодействиям. Попытка исключения любого объекта из транзакции приводит к потере важной информации о взаимосвязи объектов в транзакции.

Будем полагать, что в каждой транзакции (d, x) имеется один выделенный объект d, называемый *контейнером*, x — множество всех остальных объектов. Контейнером может быть, в зависимости от прикладной задачи, документ, клиент, операционный день, пользовательская сессия, раунд игры. Тематическая модель оценивает вероятность появления транзакции через условные распределения её объектов:

$$p(x \mid d) = \sum_{t \in T} p(t \mid d) \prod_{v \in x} p(v \mid t) = \sum_{t \in T} \theta_{td} \prod_{v \in x} \varphi_{vt}.$$

Будем также полагать, что транзакции могут иметь разные типы, отличающиеся модальностями объектов. Например, наряду с транзакциями «пользователь u кликнул объявление b на странице s» данные могут содержать транзакции «объявление b содержит слово w» и «страница s содержит слово w». Типы транзакций могут отличаться как по структуре, так и по частоте и важности.

Для построения транзакционной тематической модели ставится задача максимизации взвешенной суммы log-правдоподобий всех типов транзакций, с добавлением регуляризатора R:

$$\sum_{k \in K} \tau_k \sum_{(d,x) \in D_k} n_{kdx} \ln\Bigg(\sum_{t \in T} \theta_{td} \prod_{v \in x} \varphi_{vt}\Bigg) + R(\Phi, \Theta) \to \max_{\Phi, \Theta}; \tag{12}$$

$$\sum_{v \in W_m} \varphi_{vt} = 1,\ \varphi_{vt} \geqslant 0; \qquad \sum_{t \in T} \theta_{td} = 1,\ \theta_{td} \geqslant 0;$$

где K — множество типов транзакций, D_k — множество всех транзакций типа k, τ_k — весовые коэффициенты для балансировки типов транзакций по важности и частотности, n_{kdx} — наблюдаемые данные, число транзакций (d, x) типа k, W_m — словарь объектов модальности m.

Данная задача является, по всей видимости, одной из самых общих в тематическом моделировании [?]. Алгоритм обучения для транзакционных тематических моделей такого типа также реализован в библиотеке `BigARTM`.

1.9 От мешка слов к тематическому вниманию

Гипотеза «мешка слов» является одним из самых критикуемых предположений тематического моделирования. В рамках ARTM были разработаны пять подходов к использованию информации о порядке слов, от простого учёта *сочетаемости слов* (word co-occurrence) до тематической модели внимания.

1. *Тематическая модель контактной сочетаемости* основана на выделении часто встречающихся последовательностей из n подряд идущих термов (n-грамм). Обычно словари *n-грамм* формируются на стадии предварительной обработки текстовой коллекции и используются в качестве отдельных модальностей для каждого n, либо объединяются в одну модальность с общим словарём. Поскольку n-граммы часто оказываются терминами предметной области, они заметно улучшают интерпретируемость тем, что использовалось во многих работах [?, ?, ?, ?]. Например, триграмма «метод опорных векторов» явно указывает на тему «машинное обучение», в то время как эти же три слова по-отдельности «метод», «вектор», «опорный» имеют намного более размытую тематику.

2. *Тематическая модель дистантной сочетаемости* основана на предположении, что если два терма часто встречаются рядом, например в одном предложении или на расстоянии не более 10 термов друг от друга, то они порождаются одной общей темой. Такие пары термов называют *битермами*. Исходными данными для модели битермов служат не частоты термов в документах n_{dw}, а частоты битермов n_{vw}.

Анна Потапенко и Артём Попов в [?] предложили формировать для каждого терма v псевдо-документ, состоящий из всех термов w, образующих битермы $\{v, w\}$. К полученной коллекции псевдо-документов применяется обычный алгоритм ARTM. В задачах семантической близости слов такая модель успешно конкурирует с моделями программы `word2vec` [?] и существенно превосходит обычные тематические модели. Это означает, что в пространстве тематических векторов, как и в случае `word2vec`, выполняются тождества между ассоциативными парами слов, например,

$$\text{король} - \text{королева} = \text{мужчина} - \text{женщина};$$
$$\text{Москва} - \text{Пекин} = \text{Россия} - \text{Китай}.$$

При этом тематические векторные представления являются интерпретируемыми и разреженными. Используя кросс-энтропийные регуляризаторы, разреженность векторов удавалось доводить до 93% без потери качества модели.

В задаче семантической близости документов модель битермов уверенно опережала векторную модель DBOW [?], специально разработанную для поиска семантически близких документов.

Также в [?] была предложена мультимодальная модель битермов. Подтвердилась гипотеза, что данные о парной сочетаемости термов разных модальностей улучшают интерпретируемость тематических векторов для всех модальностей. В то же время, использование данных нетекстовых модальностей повышает качество решения задачи семантической близости слов.

Наконец, модели битермов лучше подходят для выявления тематики коллекций коротких текстов, таких как твиты или заголовки новостей. Это связано с тем, что обычным тематическим моделям не хватает данных для надёжного оценивания распределений $p(t \mid d)$ по коротким документам, в результате возникает переобучение. Модели битермов обходятся без разделения коллекции на документы — главное, чтобы для каждого терма была определена его окрестность.

В [?] предложена модель deARTM, в которой тематическая модель битермов комбинируется с иерархической моделью в рамках подхода ARTM. Показано применение модели deARTM для получения интерпретируемых тем, связанных с предпочтениями пользователей русскоязычных социальных сетей.

3. *Тематическая модель предложений* является частным случаем транзакционной модели, в которой транзакциями являются предложения. Гипотеза о том, что слова предложения порождаются одной определённой темой, не раз эксплуатировалась в литературе [?, ?]. Транзакционная тематическая модель позволяет связывать не только слова предложений, но и любые подмножества слов x, предположительно генерируемых одной общей темой из распределения $p(t \mid d)$. Это могут быть синтаксически связанные словосочетания и фразы, лексические цепочки (слова, связанные тезаурусными отношениями синонимии, «часть–целое», «общее–частное»), триплеты «объект, субъект, предикат», элементарные дискурсивные единицы.

4. *Тематическая модель с пост-обработкой E-шага.* Согласно (6), на E-шаге EM-алгоритма для каждого терма w в каждом документе d вычисляется тематиче-

ский вектор — условное вероятностное распределение $p(t \mid d, w)$. Последовательность тематических векторов термов образует матрицу $\Pi(\Phi, \Theta)$ размера $|T| \times n$, где n — суммарная длина всех документов коллекции. Столбцы в матрице Π идут в том же порядке, что и термы в исходных документах.

В [?] было предложено вводить критерии регуляризации вида $R(\Pi(\Phi, \Theta)) \to \max$, чтобы воздействовать на матрицу Π непосредственно. Например, таким способом нетрудно формализовать требование, чтобы тематические векторы термов внутри каждого предложения были сильно разрежены и схожи друг с другом, или чтобы векторы соседних предложений были схожи с вектором абзаца.

Николай Скачков в [?] использовал данный подход для повышения качества тематической сегментации документов.

Возможен и обратный подход, когда столбцы матрицы Π рассматриваются как $|T|$-мерный временной ряд, к которому применяются различные «разумные» преобразования: сглаживание, разреживание, секционирование, контрастирование и т. п. Позже было доказано [?], что любое эвристическое преобразование матрицы Π, выполненное после E-шага и перед M-шагом, эквивалентно некоторому регуляризатору, который даже не обязательно выписывать в явном виде как критерий $R(\Pi)$. Таким образом, пост-обработка E-шага оставляет большую свободу выбора алгоритмических эвристик для учёта естественного порядка слов в тексте.

5. *Быстрая линейная векторизация документов* была предложена в [?], чтобы исключить матрицу Θ из тематической модели, избежать переобучения на коротких документах и ускорить вычисления. Предлагалось определять тематический вектор документа как среднее арифметическое тематических векторов его термов:

$$\theta_{td}(\Phi) \equiv p(t \mid d) = \frac{1}{|I_d|} \sum_{i \in I_d} p(t \mid w_i), \tag{13}$$

где I_d — множество позиций, занимаемых термами документа d в сквозной нумерации термов $\{w_1, \ldots, w_n\}$ коллекции. Равенство (13) связывает параметры модели зависимостью $\Theta = \Theta(\Phi)$, фактически играя роль регуляризатора. Эксперименты на трёх текстовых коллекциях подтвердили, что при таком подходе одновременно улучшается когерентность, различность и разреженность тем [?].

Локальным контекстом терма w_i будем называть выбранную из документа подпоследовательность окружающих его термов $C_i \subset \{1, \ldots, n\}$. Контекст может быть коротким или длинным; левым, правым или двусторонним; может пропускать некоторые термы, в том числе и сам w_i; может совпадать со всем документом.

Определим тематический вектор *локального контекста* C_i по аналогии с тематическим вектором документа, усредняя с весами тематические векторы его термов:

$$\theta_{ti}(\Phi) \equiv p(t \mid C_i) = \sum_{c \in C_i} p(w_c \mid C_i)\, p(t \mid w_c),$$

где $p(w_c \mid C_i)$ — вес терма w_c в контексте C_i, называемый также *коэффициентом внимания*. Данный подход представляется наиболее перспективным и детально описан в разделе 2. Он позволяет окончательно отказаться не только от гипотезы «мешка слов», но и от деления текстовой коллекции на документы.

1.10 От подбора гиперпараметров к метаобучению

Опыт построения стратегий регуляризации, накопленный в серии работ [?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?], был собран и обобщён в библиотеке `TopicNet` [?]. Она автоматизирует перебор *гиперпараметров* (число тем, коэффициенты регуляризации, управляющие параметры алгоритма обучения), журнализацию экспериментов и визуализацию их результатов. Библиотека `TopicNet` — это высокоуровневая надстройка над `BigARTM`, которая снижает «барьеры входа» для пользователей, скрывая технические детали и многократно уменьшая объём программирования[1].

Василий Алексеев и др. в [?] использовали двухуровневую тематическую иерархию для построения «банка тем» — устойчивого набора из наиболее удачных, хорошо интерпретируемых тем. Проблема неустойчивости тематических моделей известна давно [?], но на практике зачастую игнорируется. Неустойчивость проявляется, в частности, в том, что при повторных запусках EM-алгоритма по одним и тем же данным, но из различных начальных приближений или при различной рандомизации, получаются разные наборы тем. Поэтому в [?] предлагается отбирать после каждого запуска наилучшие темы и откладывать их в «банк тем». При построении следующей модели «банк тем» используется в качестве набора родительских тем верхнего уровня. Альтернативный подход, развивающий эту идею, предлагается в [?]. Более ранние подходы к построению полного набора тем, основанные на построении выпуклой оболочки [?, ?] вычислительно более ресурсоёмкие.

Построение «банка тем», хотя и поддерживается библиотекой `TopicNet`, может оказаться слишком ресурсоёмким для отдельных практических задач. Поэтому в [?, ?] ставится задача *мета-обучения* (meta-learning) — построив «банк тем» для различных текстовых коллекций, подобрать универсальную стратегию регуляризации, ведущую к построению набора тем, близкого к «банку тем», но за один запуск. На данный момент эта задача ещё далека от своего окончательного решения.

Другой подход к подбору гиперпараметров для `BigARTM` был предложен в [?]. Он основан на эволюционном алгоритме для построения стратегии адаптивного многоэтапного изменения гиперпараметров. В последующей работе [?] этот метод был дополнен суррогатной моделью для оценивания качества модели, что позволило примерно в 1,5 раза сократить время поиска гиперпараметров. Использовался обычный для *автоматического машинного обучения* (AutoML) подход: для каждой точки в пространстве гиперпараметров модель обучается заново по обучающей выборке, затем её качество оценивается по тестовой выборке.

На больших данных такая стратегия представляется слишком ресурсоёмкой. Корректировать гиперпараметры можно было бы и чаще — обработав очередную порцию данных, не дожидаясь сходимости, оценивать текущее качество моделей приближённо, но быстро, на небольших случайных контрольных подвыборках, по многим метрикам качества. Построение такого рода быстрых многокритериальных стратегий регуляризации пока остаётся открытой проблемой.

[1] `https://machine-intelligence-laboratory.github.io/TopicNet` — документация и открытый исходный код библиотеки TopicNet. Разработка лаборатории машинного интеллекта МФТИ.

1.11 От открытых проблем к приоритетам исследований

В области вероятностного тематического моделирования накопилось некоторое количество проблем, не решённых в рамках байесовского подхода, но имеющих перспективы найти свои окончательные решения в ARTM.

1. Остаётся открытой проблема перехода от гипотезы «мешка слов» к модели связного текста, достаточно удобной, простой и расширяемой, чтобы стать стандартом де-факто для многочисленных модификаций, обобщений и приложений, заменив наконец в этой роли морально устаревшую модель LDA.

Создание тематической модели внимания с поддержкой аддитивной регуляризации, мультимодальных и транзакционных данных, представляется наиболее перспективным подходом.

Это потребует в самом ближайшем будущем новых исследований для проверки гипотез о регуляризирующем воздействии внимания $\theta_{ti}(\Phi)$ на устойчивость, когерентность, разреженность и другие характеристики качества тематической модели.

2. Остаётся открытой проблема плохой интерпретируемости тем из-за тематической несбалансированности данных. Если объёмы тем в исходной коллекции различаются в сотни раз, то крупные темы дробятся на темы-дубликаты, мелкие объединяются и образуют мусорные темы [?]. Причина этого явления кроется в том, что для максимизации правдоподобия в задаче матричного разложения выгодно выравнивать темы по объёму, поэтому распределение $p(t)$ имеет тенденцию к выглаживанию. Ранее предлагалось вводить распределение $p(t)$ в модель явным образом, воздействуя либо на столбцы матрицы Θ [?], либо на строки матрицы Φ [?]. Эффекты удаления мусорных и линейно зависимых тем наблюдались также при использовании регуляризатора разреживания распределения $p(t)$ в [?] и регуляризатора декоррелирования.

Эвристические способы перевзвешивания тем, аналогичные применяемым в задачах классификации с несбалансированными классами, не ведут к желаемому результату, как было доказано в [?]. Проблема несбалансированности в тематических моделях, по всей видимости, является более трудной, чем в задачах классификации.

Есть основания полагать, что тематические модели внимания лучше подходят для анализа несбалансированных коллекций, поскольку в них нет матрицы Θ, наиболее подверженной этому эффекту. Как и в предыдущем пункте, это потребует проверки гипотез о способности модели внимания к выявлению тематического дисбаланса в данных и устранению эффектов дробления и слияния тем.

3. Проблема автоматического оценивания интерпретируемости тем считалась закрытой циклом работ [?, ?, ?], где был выработан критерий *когерентности* темы (coherence), наиболее коррелирующий с экспертными оценками интерпретируемости. Позже было замечено, что топовые слова тем, используемые в оценках когерентности, слишком разреженно представлены в текстовой коллекции, занимая в совокупности не более 1–2% от её объёма. Другими словами, когерентность недостаточно репрезентативна [?].

Василий Алексеев в [?] предложил оценивать *внутритекстовую когерентность*, собирая по тексту статистику тематической близости для всех битермов. Внутритекстовая когерентность представляется наиболее адекватной величиной для автоматического оценивания интерпретируемости тем в тематических моделях внимания,

однако проверка этой гипотезы потребует новых калибровочных экспериментов, аналогичных [?, ?, ?].

4. Вопросы токенизации текстовой коллекции, в том числе выделения терминов, словосочетаний, названий, аббревиатур, обычно выносятся на стадию предварительной обработки данных. Известно, что термины и словосочетания существенно повышают интерпретируемость тем. Проблема в том, что словарь n-грамм слишком быстро растёт с ростом коллекции, и для сохранения разумного размера модели требуется отбор тематически значимых словосочетаний. Задача оптимального сокращения словаря (vocabulary reduction) обычно решается на стадии постобработки и требует многократного перестроения модели. Открытой проблемой остаётся встраивание эффективных методов отбора словосочетаний в алгоритм тематического моделирования. Для пакетных и онлайновых версий EM-алгоритма эта задача не имеет общепринятого решения.

5. Остаётся открытой проблема автоматического краткого изложения (суммаризации) тем. Большинство приложений тематического моделирования связано с разведочным анализом текстовой коллекции и требует информативной но лаконичной визуализации тем в пользовательском интерфейсе. Каждая тема должна «уметь рассказать о себе». Задача автоматического именования тем была впервые поставлена и решена в [?]. Однако задача суммаризации тем в литературе даже не ставилась вплоть до появления больших языковых моделей.

Суммаризация тем с использованием больших языковых моделей требует аккуратного предварительного отбора и ранжирования тематически релевантных словосочетаний, фраз, предложений. Данная задача на сегодняшний день также не имеет общепринятого (стандартного де-факто) решения.

6. Остаётся открытой проблема автоматического добавления новых тем и удаления старых в текстовых потоках. Точнее, задача обнаружения и отслеживания тем (topic detection and tracking, TDT) хорошо известна в литературе, начиная с отчёта DARPA [?], и имеет множество решений, причём не только на основе тематического моделирования. Однако среди них нет универсальных, в которых эффективно решался бы целостный комплекс подзадач: своевременное и статистически надёжное обнаружение новой темы, оптимальный выбор наименее актуальных тем для удаления, пополнение словаря новыми термами, архивация старых термов и тем для поддержания примерно постоянного числа тем в модели. Существующие решения либо игнорируют часть из этих подзадач, либо узко специализированы для анализа конкретного типа данных (новости, патентные базы, научные публикации, социальные медиа, электронная почта).

7. Остаётся открытой проблема автоматического подбора гиперпараметров, особенно в режиме онлайновой пакетной обработки коллекции. При использовании нескольких регуляризаторов и модальностей их весовые коэффициенты обычно подбираются вручную, путём многократного перестроения модели по всей коллекции. В некоторых приложениях приходится перебирать число тем, для иерархических моделей — на каждом уровне. Иногда подбираются также параметры EM-алгоритма: число итераций, размер пакетов, темп забывания предыдущих пакетов, число пакетов для обновления матрицы Ф. Применение генетических алгоритмов или методов AutoML проводит к громоздким и вычислительно трудоёмким алгоритмам. Представляется перспективным многокритериальное оценивание качества модели по пакетам с привлечением методов планирования экспериментов.

2 Тематическая модель локальных контекстов

В данном разделе предлагается новое семейство тематических моделей локальных контекстов, свободное от двух наиболее сильных и часто критикуемых гипотез вероятностного тематического моделирования. Первая — гипотеза «мешка слов», согласно которой порядок слов не имеет значения для определения тематики документа. Вторая — гипотеза тематической однородности документов, согласно которой тематика слова в документе определяется тематикой всего документа, а не локальным контекстом слова.

В предлагаемом далее семействе моделей вообще не предполагается разделения текстовой коллекции на документы. Вместо этого для каждого слова определяется его локальный контекст. Мы оставляем значительную свободу выбора в определении контекста: он может быть коротким или длинным; левым, правым или двусторонним; непрерывным или исключающим некоторые слова; содержащим центральное слово или исключающим его. Более того, мы вводим коэффициенты внимания, определяющие важность слов из контекста. Каждое слово получает не только контекстно-независимое тематическое векторное представление (эмбединг), но и контекстно-зависимый тематический эмбединг в каждом контексте употребления.

Перечисленные обстоятельства сближают предлагаемое семейство моделей, которые мы называем *внимательными аддитивно регуляризованными тематическими моделями* (Attentive Additively Regularized Topic Model, AARTM), с глубокими нейросетевыми моделями языка — кодировщиками типа BERT [**?**]. В то же время алгоритм обучения AARTM остаётся, по сути, быстрым методом мягкой кластеризации текстовой коллекции, не требующим ресурсоёмких вычислений.

Мы показываем, что учёт локальных контекстов в моделях семейства AARTM приводят к регуляризирующему эффекту, улучшающему когерентность, следовательно, и интерпретируемость тем [**?**, **?**, **?**]. Кроме того, в модель вводится распределение $p(t)$ — тематический эмбединг текстовой коллекции, позволяющий оказывать управляющие воздействия на балансировку тем в процессе обучения модели. Тем самым закладывается теоретическая основа для решения двух фундаментальных проблем тематического моделирования — проблемы дефицита интерпретируемости тем и проблемы тематической несбалансированности данных.

2.1 Тематические эмбединги

Тематическими эмбедингами будем называть дискретные вероятностные распределения вида $p(t \mid x)$ на множестве тем T.

Описанная в разделе 1.3 модель является несимметричной в том смысле, что взаимосвязь документов с темами описывается тематическими эмбедингами $p(t \mid d)$, тогда как взаимосвязь термов с темами описывается частотными словарями тем $p(w \mid t)$. Результатом E-шага также является тематический эмбединг $p(t \mid d, w)$ терма w в контексте документа d. Кроме того, в EM-алгоритме естественным образом возникает несмещённая оценка тематического эмбединга всей коллекции:

$$p(t) = \underset{t \in T}{\operatorname{norm}} \left(\sum_{d \in D} \sum_{w \in d} n_{dw} p_{tdw} \right) = \frac{1}{n} \sum_{d \in D} n_{dt} = \frac{1}{n} \sum_{w \in W} n_{tw} = \frac{n_t}{n},$$

где n_t — оценка числа термов темы t во всей коллекции, n — суммарная длина всех документов коллекции.

Нормировка столбцов матрицы Φ может быть не удобна в онлайновых алгоритмах, когда словарь пополняется новыми термами при обработке новых документов. В таком случае была бы удобнее нормировка по строкам, когда вместо частотных словарей $p(w\,|\,t)$ оцениваются тематические эмбединги термов $p(t\,|\,w)$.

Чтобы из обычной тематической модели (1) получить симметризованную, в которой все параметры являются тематическими эмбедингами, воспользуемся тождеством $p(t\,|\,w)p(w) = p(w\,|\,t)p(t)$:

$$p(w\,|\,d) = \sum_{t\in T} \frac{p(t\,|\,w)p(w)}{p(t)} p(t\,|\,d) = \sum_{t\in T} \varphi_{tw}\theta_{td}\frac{p(w)}{p(t)}. \tag{14}$$

Для оценивания параметров модели $\Phi = (\varphi_{tw})_{T\times W}$ и $\Theta = (\theta_{td})_{T\times D}$ будем максимизировать функционал регуляризированного логарифма правдоподобия:

$$L(\Phi,\Theta) = \sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{t\in T} \varphi_{tw}\theta_{td}\frac{p(w)}{p(t)} + R(\Phi,\Theta) \;\to\; \max_{\Phi,\Theta} \tag{15}$$

при ограничениях неотрицательности и нормировки:

$$\sum_{t\in T} \varphi_{tw} = 1, \quad \varphi_{tw} \geqslant 0; \qquad \sum_{t\in T} \theta_{td} = 1, \quad \theta_{td} \geqslant 0. \tag{16}$$

Теорема 2. *Пусть функция $R(\Phi,\Theta)$ непрерывно дифференцируема. Точка (Φ,Θ) локального экстремума задачи (15), (16) удовлетворяет следующей системе уравнений со вспомогательными переменными $p_{tdw} = p(t\,|\,d,w)$, если из решения исключить нулевые столбцы матриц Φ и Θ:*

$$p_{tdw} = \underset{t\in T}{\operatorname{norm}}\big(\varphi_{tw}\theta_{td}/p(t)\big); \qquad p(t) = \frac{1}{n}\sum_{d\in D}\sum_{w\in d} n_{dw}p_{tdw}; \tag{17}$$

$$\varphi_{tw} = \underset{t\in T}{\operatorname{norm}}\left(n_{tw} + \varphi_{tw}\frac{\partial R}{\partial \varphi_{tw}}\right); \qquad n_{tw} = \sum_{d\in D} n_{dw}p_{tdw}; \tag{18}$$

$$\theta_{td} = \underset{t\in T}{\operatorname{norm}}\left(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right); \qquad n_{td} = \sum_{w\in d} n_{dw}p_{tdw}. \tag{19}$$

Доказательство строится на применении Леммы 1 к задаче (15), (16).
Согласно лемме, необходимо найти следующие частные производные:

$$\varphi_{tw}\frac{\partial L}{\partial \varphi_{tw}} = \varphi_{tw}\sum_{d\in D} n_{dw}\frac{\theta_{td}p(w)}{p(w\,|\,d)\,p(t)} + \varphi_{tw}\frac{\partial R}{\partial \varphi_{tw}} = \sum_{d\in D} n_{dw}p_{tdw} + \varphi_{tw}\frac{\partial R}{\partial \varphi_{tw}};$$

$$\theta_{td}\frac{\partial L}{\partial \theta_{td}} = \theta_{td}\sum_{w\in d} n_{dw}\frac{\varphi_{tw}p(w)}{p(w\,|\,d)\,p(t)} + \theta_{td}\frac{\partial R}{\partial \theta_{td}} = \sum_{w\in d} n_{dw}p_{tdw} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}.$$

Теорема доказана.

Утверждение теоремы справедливо для произвольного распределения $p(t)$. Более корректной, казалось бы, должна быть подстановка $p(t) = \sum_w \varphi_{tw}p(w)$, однако это привело бы к усложнению доказательства. Мы сразу подставляем несмещённую оценку $p(t)$, поскольку при практической реализации в правые части формул М-шага

также подставляются несмещённые оценки $\varphi^*_{tw} = \text{norm}_t(n_{tw})$ и $\theta^*_{td} = \text{norm}_t(n_{td})$, согласованные с несмещённой оценкой $p(t)$.

Подчеркнём отличия ЕМ-алгоритма (17)–(19) от обычной версии (6)–(8).

1. Матрица Φ поменяла ориентацию, теперь столбцы соответствуют термам и являются тематическими эмбедингами.

2. Формула М-шага (18) для матрицы Φ осталась почти той же; смысл счётчиков n_{tw} не изменился; единственное отличие в том, что нормировка по $w \in W$ заменена нормировкой по $t \in T$.

3. Формула М-шага (19) для матрицы Θ не изменилась.

4. Распределение $p(t)$ даёт возможность управлять балансировкой тем в случае тематически несбалансированных коллекций [?], добиваясь наилучшей интерпретируемости каждой темы независимо от её объёма.

2.2 Локальные контексты и внимание

Предположение о том, что тематический эмбединг документа $p(t \mid d)$ получается усреднением тематических эмбедингов его термов $p(t \mid w)$, $w \in d$, было впервые введено в [?]. Это предположение фактически играет роль регуляризатора, поскольку в оптимизационной задаче появляется ограничение-равенство вида $\Theta = f(\Phi)$. Такой подход позволяет убрать матрицу Θ из модели, а вычисление вектора $p(t \mid d)$ для любого документа d производить быстро — без итерационного процесса, за один проход по документу. Эксперименты в [?] показали, что сокращение размерности пространства параметров снижает переобучение и повышает качество модели.

Дальнейшим развитием данной идеи является предположение о том, что тематический эмбединг произвольного фрагмента текста также получается усреднением тематических эмбедингов составляющих его термов [?]. Это позволяет отказаться от двух наиболее ограничивающих гипотез тематического моделирования — во-первых, от гипотезы «мешка слов»; во-вторых, от гипотезы, что тематика слова определяется всем документом. Следование второй гипотезе может приводить к переобучению в случае коротких документов и недообучению на длинных документах с тематически неоднородной сегментной структурой.

Обозначим через $\{w_1, \ldots, w_n\}$ конкатенацию всех документов текстовой коллекции, в естественном порядке их термов, где n — суммарная длина коллекции.

Обозначим через $C_i \subset \{1, \ldots, n\}$ позиции термов, образующих контекст (локальную окрестность) терма w_i. Контекст можно определять по-разному: он может быть коротким или длинным; левым, правым или двусторонним; непрерывным или исключающим некоторые термы; содержащим терм w_i или исключающим его. Контекст не может выходить за пределы документа, захватывая термы других документов.

Вероятностная тематическая модель локального контекста предсказывает появление терма в i-й позиции по его контексту:

$$p(w \mid C_i) = \sum_{t \in T} p(w \mid t) p(t \mid C_i) = \sum_{t \in T} p(t \mid w) \frac{p(w)}{p(t)} p(t \mid C_i).$$

Можно заметить, что $p(w \mid C_i)$ получается из обычной тематической модели $p(w \mid d)$ формальной заменой документа d на контекст C_i.

Введём тематический эмбединг контекста как средневзвешенное тематических эмбедингов термов:

$$p(t \mid C_i) \equiv \theta_{ti} = \sum_{c \in C_i} \alpha_{ci} p(t \mid w_c), \qquad \sum_{c \in C_i} \alpha_{ci} = 1, \qquad \alpha_{ci} \geqslant 0,$$

где α_{ci} — *коэффициенты внимания*, показывающие, насколько важен (с каким весом учитывается) терм w_c из контекста C_i для терма w в позиции i.

С учётом симметризации модель локального контекста принимает вид

$$p(w \mid C_i) = \sum_{t \in T} \varphi_{tw} \frac{p(w)}{p(t)} \sum_{c \in C_i} \alpha_{ci} \varphi_{tw_c}.$$

Для оценивания параметров модели $\Phi = (\varphi_{tw})_{T \times W}$ будем максимизировать функционал регуляризированного логарифма правдоподобия:

$$L(\Phi) = \sum_{i=1}^{n} \ln p(w_i \mid C_i) + R(\Phi) \;\to\; \max_{\Phi}. \tag{20}$$

при ограничениях неотрицательности и нормировки:

$$\sum_{t \in T} \varphi_{tw} = 1, \quad \varphi_{tw} \geqslant 0. \tag{21}$$

Теорема 3. *Пусть функция $R(\Phi)$ непрерывно дифференцируема. Точка (Φ) локального экстремума задачи (20), (21) удовлетворяет системе уравнений со вспомогательными переменными $p_{ti} = p(t \mid C_i, w_i)$, $\theta_{ti} = p(t \mid C_i)$, N_{tw}, q_{wi}, если из решения исключить нулевые столбцы матрицы Φ:*

$$p_{ti} = \underset{t \in T}{\operatorname{norm}}\big(\varphi_{tw_i} \theta_{ti} / p(t)\big); \qquad \theta_{ti} = \sum_{c \in C_i} \alpha_{ci} \varphi_{tw_c} \tag{22}$$

$$N_{tw} = \sum_{i=1}^{n} q_{wi} \frac{p_{ti}}{\theta_{ti}}; \qquad q_{wi} = \sum_{c \in C_i} \alpha_{ci} [w_c = w]; \tag{23}$$

$$\varphi_{tw} = \underset{t \in T}{\operatorname{norm}}\left(n_{tw} + \varphi_{tw} N_{tw} + \varphi_{tw} \frac{\partial R}{\partial \varphi_{tw}} \right); \qquad n_{tw} = \sum_{i=1}^{n} p_{ti} [w_i = w]; \tag{24}$$

$$p(t) = \frac{1}{n} \sum_{i=1}^{n} p_{ti}. \tag{25}$$

Доказательство строится на применении Леммы 1 к задаче (20), (21).

Запишем максимизируемый функционал, заменив индекс t на s:

$$L(\Phi) = \sum_{i=1}^{n} \ln\left(\sum_{s \in T} \varphi_{sw_i} \frac{p(w_i)}{p(s)} \theta_{si} \right) + R(\Phi) \;\to\; \max_{\Phi}$$

Найдём сначала две вспомогательные частные производные:

$$\frac{\partial \varphi_{sw_i}}{\partial \varphi_{tw}} = [w_i = w][t = s];$$

$$\frac{\partial \theta_{si}}{\partial \varphi_{tw}} = \frac{\partial}{\partial \varphi_{tw}} \left(\sum_{c \in C_i} \alpha_{ci} \varphi_{sw_c} \right) = \sum_{c \in C_i} \alpha_{ci} [w_c = w][t = s] = q_{wi} [t = s].$$

Согласно лемме, необходимо найти частные производные:

$$\begin{aligned}
\varphi_{tw}\frac{\partial L}{\partial \varphi_{tw}} &= \sum_{i=1}^{n}\frac{p(w_i)}{p(w_i\,|\,C_i)}\sum_{s\in T}\frac{1}{p(s)}\varphi_{tw}\frac{\partial}{\partial\varphi_{tw}}\big(\varphi_{sw_i}\theta_{si}\big) + \varphi_{tw}\frac{\partial R}{\partial\varphi_{tw}} = \\
&= \sum_{i=1}^{n}\frac{p(w_i)}{p(w_i\,|\,C_i)\,p(t)}\varphi_{tw}\big(\theta_{ti}[w_i=w] + \varphi_{tw_i}q_{wi}\big) + \varphi_{tw}\frac{\partial R}{\partial\varphi_{tw}} = \\
&= \sum_{i=1}^{n}\frac{\varphi_{tw_i}p(w_i)\theta_{ti}}{p(w_i\,|\,C_i)\,p(t)}\left([w_i=w] + \varphi_{tw}\frac{q_{wi}}{\theta_{ti}}\right) + \varphi_{tw}\frac{\partial R}{\partial\varphi_{tw}} = \\
&= \sum_{i=1}^{n}p_{ti}[w_i=w] + \varphi_{tw}\sum_{i=1}^{n}q_{wi}\frac{p_{ti}}{\theta_{ti}} + \varphi_{tw}\frac{\partial R}{\partial\varphi_{tw}} = \\
&= n_{tw} + \varphi_{tw}N_{tw} + \varphi_{tw}\frac{\partial R}{\partial\varphi_{tw}};
\end{aligned}$$

Собирая воедино все выведенные уравнения, получаем утверждение теоремы.
Теорема доказана.

Утверждение теоремы справедливо для произвольного распределения $p(t)$, но мы сразу подставляем несмещённую оценку (25), рассчитывая, что при практической реализации в (24) также подставляются несмещённые оценки $\varphi^*_{tw} = \mathrm{norm}_t(n_{tw})$, согласованные с несмещённой оценкой $p(t)$.

Интерпретации вспомогательных переменных:

n_{tw} — сколько раз терм w относится к теме t во всей коллекции;

q_{wi} — вес терма $w \in W$ в контексте C_i (равен нулю, если терма в контексте нет);

N_{tw} — суммарный вес терма w в контекстах, относящихся к теме t.

Формула M-шага (24) означает, что контекст играет роль регуляризатора. Добавка $\varphi_{tw}N_{tw}$ всегда неотрицательна и увеличивает условную вероятность $p(t\,|\,w)$, если терм w часто наблюдается в таких локальных контекстах C_i, где терм w_i с большой вероятностью относится к теме t.

С другой стороны, регуляризирующая добавка может оказаться пренебрежимо малой, $\varphi_{tw}N_{tw} \ll n_{tw}$. Если она мала всегда, то нет необходимости вычислять вспомогательные переменные q_{wi} и N_{tw}, значит, алгоритм можно упростить и ускорить.

Регуляризирующее воздействие локальных контекстов аналогично тематическим моделям дистрибутивной семантики, таким, как BTM (biterm topic model) [?], WTM (word topic model) [?], WNTM (word network topic model) [?], их аддитивно регуляризованные версии [?], а также регуляризаторам когерентности [?, ?]. Регуляризирующая добавка $\varphi_{tw}N_{tw}$ способствует образованию тем из термов, часто встречающихся рядом в общем контексте, следовательно, повышает когерентность и интерпретируемость тем. Аналогичным образом производится сближение эмбедингов термов, часто встречающихся в схожих контекстах, в языковых моделях семейства word2vec и их многочисленных обобщениях [?, ?, ?, ?].

С третьей стороны, регуляризирующая добавка $\varphi_{tw}N_{tw}$ интересна тем, что она естественным образом определяет тематику терма w по его локальным контекстам, если он является редким в коллекции или вообще встречается впервые. Такой встроенной возможности никогда не было в классических тематических моделях на основе PLSA или LDA.

Таким образом, мы определили семейство моделей локального контекста, оставив большую свободу выбора как способов определения контекста C_i и коэффициентов внимания α_{ci}, так и эвристических модификаций EM-алгоритма. Поиск наилучшего сочетания этих способов и модификаций является предметом дальнейших эмпирических исследований.

2.3 Аналогия с нейросетевыми моделями внимания

Вычисление контекстного тематического вектора $p(t \mid w_i, C_i)$ по бесконтекстному тематическому вектору $p(t \mid w_i)$ того же терма w_i происходит во многом аналогично свёрточной нейросетевой модели GCNN (gated convolutional neural network) [?]. Некоторая аналогия прослеживается также с нейросетевой моделью внимания query–key–value [?]. Здесь также суммируются векторы из контекста $p(t \mid w_c)$, $c \in C_i$, называемые *значениями* (value), которые также умножаются на нормированные весовые коэффициенты α_{ci}, которые также показывают, насколько близки вектор терма w_c из контекста, называемый *ключом* (key) и вектор терма w_i, называемый *запросом* (query).

Но есть и существенные отличия, их тоже три.

Во-первых, тематические векторы неотрицательные, нормированные, интерпретируемые и, как правило, разреженные, в то время как нейросетевые векторные представления (эмбединги) не ограниченные, не интерпретируемые и плотные.

Во-вторых, в модели внимания векторы запросов, ключей и значений умножаются на матрицы линейных преобразований, оптимизируемые при обучении сети. Такая параметризация существенно обогащает модель, однако наша тематическая модель радикально проще и не имеет обучаемых параметров.

В-третьих, в нашей модели параметры α_{ci} могут выражать лишь позиционную близость термов w_c и w_i, в то время как модель внимания вычисляет α_{ci} через семантическую близость эмбедингов w_c и w_i. Отчасти это компенсируется тем, что тематические векторы-ключи $p(t \mid w_c)$ не просто складываются с весами α_{ci}, но сначала домножаются покомпонентно на вектор-запрос $p(t \mid w_i)$, который играет роль фильтра, пропуская только те темы, которые есть у терма w_i. В итоге сумма формируется преимущественно из векторов-значений $p(t \mid w_c)$ термов w_c, тематически близких вектору-запросу w_i, что в значительной степени аналогично как модели внимания, так и свёрточной модели GCNN, но без матриц параметров.

Таким образом, тематическая модель локальных контекстов является предельно упрощённым аналогом нейросетевой модели внимания query–key–value, но не имеет обучаемых параметров. Развитие и анализ качества тематических моделей внимания является целью ближайших исследований, включая вопросы оптимизации коэффициентов внимания α_{ci} и дополнительной параметризации модели.

2.4 Определение коэффициентов внимания

Введение коэффициентов внимания α_{ci} позволяет сделать границы контекстов нечёткими, уменьшая α_{ci} по мере увеличения дистанции между термами w_i и w_c.

Рассмотрим простой и вычислительно эффективный способ задания весов с помощью двух *экспоненциальных скользящих средних* (ЭСС), которые усредняют тематические эмбединги термов по рекуррентным формулам, пробегая по тексту в двух

направлениях — слева направо и справа налево:

$$\vec{p}(t\,|\,i) = \gamma_i \cdot \varphi_{tw_i} + (1-\gamma_i) \cdot \vec{p}(t\,|\,i-1), \quad i = 1,\ldots,n, \quad \gamma_1 = 1;$$
$$\overleftarrow{p}(t\,|\,i) = \gamma_i \cdot \varphi_{tw_i} + (1-\gamma_i) \cdot \overleftarrow{p}(t\,|\,i+1), \quad i = n,\ldots,1, \quad \gamma_n = 1.$$

Для выбора коэффициента сглаживания $\gamma_i \in [0,1]$ можно ориентироваться на известную для ЭСС оценку $\gamma_i \approx \frac{1}{h}$, где h — число усредняемых векторов. Например, $\gamma_i = 0.1$ приблизительно соответствует усреднению 10 векторов.

Если $\gamma_i = \gamma$ — константа, то экспоненциальное скользящее среднее определяет веса, убывающие с расстоянием между позициями $|i-c|$ по геометрической прогрессии: $\alpha_{ci} = \gamma(1-\gamma)^{|i-c|}$.

Коэффициент γ_i можно менять в зависимости от позиции i. Его увеличение вплоть до $\gamma_i = 1$ позволяет «забыть» накопленную к данному моменту информацию о контексте, например, при переходе к следующему документу или к следующей секции документа. Уменьшение γ_i вплоть до нуля позволяет игнорировать терм, например, если это стоп-слово или слово общей лексики.

Зная тематические эмбединги левого и правого контекста $\vec{p}(t\,|\,i)$ и $\overleftarrow{p}(t\,|\,i)$ для каждой позиции i, можно отвечать на многие вопросы о тематической структуре текста и любых его фрагментов.

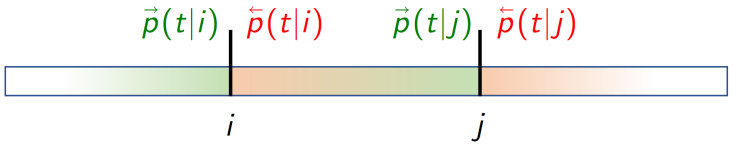
Тематический эмбединг двустороннего контекста определяется как среднее арифметическое или (для общности) среднее взвешенное левого и правого контекста:

$$\theta_{ti} \equiv p(t\,|\,C_i) = \beta\vec{p}(t\,|\,i) + (1-\beta)\overleftarrow{p}(t\,|\,i).$$

Границу i между тематическими сегментами длинного текстового документа можно определять по максимуму различия между эмбедингами $\vec{p}(t\,|\,i)$ и $\overleftarrow{p}(t\,|\,i)$.

Анализ отклонений коротких и длинных ЭСС, соответственно при больших и малых γ_i, позволяет определить, содержит ли текст вкрапления тематически разнородных фрагментов, и где они находятся.

Рассмотрим *двунаправленные тематические эмбединги* для пары термов w_i и w_j:



Тематику короткого сегмента $[i\,..\,j]$ можно оценить, усредняя эмбединги правого контекста его начала i и левого контекста его конца j:

$$p(t\,|\,i\,..\,j) = \tfrac{1}{2}\big(\overleftarrow{p}(t\,|\,i) + \vec{p}(t\,|\,j)\big).$$

Тематическая однородность сегмента $[i\,..\,j]$ оценивается тем, насколько схожи эмбединги $\overleftarrow{p}(t\,|\,i)$ и $\vec{p}(t\,|\,j)$.

2.5 EM-алгоритм для модели локального контекста

На входе алгоритму передаётся текстовая коллекция, задаётся число итераций по всей коллекции K и число итераций по каждому документу L.

Алгоритм 1. EM-алгоритм для построения модели Attentive ARTM.

вход: текстовая коллекция, число тем $|T|$, параметры K, L, β, $\vec{\gamma}_i$, $\overleftarrow{\gamma}_i$;
выход: матрица Φ, векторы термов документов p_{ti};

1 инициализация $\varphi_{tw} := \operatorname{norm}_t(\text{rand})$; $N_{tw} := 0$; $n_t := 1$ для всех $w \in W$, $t \in T$;
2 **для всех** итераций $k = 1..K$ (проходов по всей коллекции)
3 инициализация: $(n_{tw}, N_{tw}, \tilde{n}_t) := 0$; для всех $w \in W$, $t \in T$;
4 **для всех** документов $d \in D$
5 $p_{ti} := \varphi_{tw_i}$ для всех $t \in T$, $i = 1..n_d$;
6 **для всех** $l = 1..L$ (аналог L блоков внимания в трансформере)
7 $\theta_{ti} := \mathsf{BidirEMA}\big([p_{ti}\colon t \in T,\ i = 1..n_d]\big)$;
8 $p_{ti} := \underset{t\in T}{\operatorname{norm}}\big(p_{ti}\theta_{ti}/n_t\big)$ для всех $t \in T$, $i = 1..n_d$;
9 $q_{wi} := \mathsf{BidirEMA}\big([[w_i{=}w]\colon w \in d,\ i = 1..n_d]\big)$;
10 $N_{tw} := N_{tw} + q_{wi}p_{ti}/\theta_{ti}$ для всех $t \in T$, $w \in d$, $i = 1..n_d$;
11 $n_{tw_i} := n_{tw_i} + p_{ti}$; $\tilde{n}_t := \tilde{n}_t + p_{ti}$ для всех $t \in T$, $i = 1..n_d$;
12 $\varphi_{tw} := \underset{t\in T}{\operatorname{norm}}\Big(n_{tw} + \frac{n_{tw}}{n_w}N_{tw} + \frac{n_{tw}}{n_w}\frac{\partial R}{\partial \varphi_{tw}}\Big|_{\varphi_{tw}=\frac{n_{tw}}{n_w}}\Big)$ для всех $t \in T$, $w \in W$;
13 $n_t := \tilde{n}_t$ для всех $t \in T$;

14 **функция** BidirEMA: двунаправленное экспоненциальное скользящее среднее:
  вход: $(x_{hi}\colon h \in H,\ i = 1..n)$ — последовательность n H-мерных векторов;
  выход: $(y_{hi}\colon h \in H,\ i = 1..n)$ — последовательность n H-мерных векторов;
15 $\vec{y}_{hi} := \vec{\gamma}_i\, x_{hi} + (1{-}\vec{\gamma}_i)\, \vec{y}_{h,i-1}$ для всех $h \in H$, $i = 1..n$, $\vec{\gamma}_1 = 1$;
16 $\overleftarrow{y}_{hi} := \overleftarrow{\gamma}_i\, x_{hi} + (1{-}\overleftarrow{\gamma}_i)\, \overleftarrow{y}_{h,i+1}$ для всех $h \in H$, $i = 1..n$, $\overleftarrow{\gamma}_n = 1$;
17 $y_{hi} := \beta\vec{y}_{hi} + (1{-}\beta)\overleftarrow{y}_{hi}$ для всех $h \in H$, $i = 1..n$;

Внешний цикл по k от 1 до K — это итерации EM-алгоритма. Каждая итерация — это проход по всем документам коллекции. Разбиение на документы является чисто техническим и связано только с удобством хранения и загрузки данных, никакие параметры модели не оцениваются для документов.

Внутренний цикл по l от 1 до L — это итерации по каждому документу. Эти итерации не обязательны, теоретически достаточно одной. Они введены как эвристическое обобщение алгоритма, имеющее некоторую аналогию с тем, как в больших языковых моделях L моделей внимания, обрабатывающих последовательность эмбедингов, образуют архитектуру трансформера [**?**].

Шаг 5: перед первой итерацией по документу в массив p_{ti} записывается последовательность бесконтекстных эмбедингов термов, $p_{ti} = p(t \,|\, w_i)$, $i = 1, \ldots, n_d$.

Шаг 7: функция BidirEMA преобразует эту последовательность в тематические эмбединги контекста $\theta_{ti} = p(t \,|\, C_i)$, $i = 1, \ldots, n_d$, вычисляя коэффициенты внимания неявно, с помощью двух экспоненциальных скользящих средних, направленных по документу слева направо и справа налево.

Шаг 8: итерация по документу завершается вычислением контекстных эмбедингов термов $p_{ti} = p(t \,|\, C_i, w_i)$ по формулам E-шага (22), которые записываются в тот же массив p_{ti}. На следующем цикле по l контекстные эмбединги обрабатываются вместо бесконтекстных, по аналогии с архитектурой трансформера.

Шаг 9: функция BidirEMA используется ещё раз для вычисления последовательности векторов (q_{wi}), $i = 1, \ldots, n_d$ согласно формуле (23). Размерность этих векторов равна числу различных термов в документе, $w \in d$.

Шаг 10: в массиве N_{tw} накапливается статистика сочетаемости терма w с контекстами темы t по всей коллекции, согласно формуле (23). Заметим, что это наиболее трудоёмкий шаг в данном алгоритме, имеющий вычислительную сложность $O(n_d^2\,|T|)$ для каждого документа d.

Шаг 11: в массиве n_{tw} накапливается статистика сочетаемости терма w с темой t по всей коллекции, согласно формуле (24).

Шаг 12: обновление матрицы параметров Φ производится в конце каждого прохода по коллекции, по формуле M-шага (24). Вместо φ_{tw} подставляются несмещённые оценки $\frac{n_{tw}}{n_w}$ для улучшения сходимости [?].

Для больших коллекций обновления матрицы Φ можно производить чаще, по мере накопления достаточного объёма статистики в счётчиках n_{wt}, что ускорит сходимость итерационного процесса [?, ?, ?]. Если коллекция настолько большая, что матрица Φ успевает сойтись до окончания первой итерации, то к такой коллекции применима онлайновая модификация EM-алгоритма.