

Аддитивная регуляризация тематических моделей: теория и приложения

Константин Вячеславович Воронцов

(МГУ, МФТИ, ФИЦ ИУ РАН)

Аннотация

Данный обзор отражает состояние исследований по вероятностному тематическому моделированию (PTM) в рамках подхода аддитивной регуляризации (additive regularization for topic modeling, ARTM). В отличие от байесовского обучения, наиболее широко принятого в данной области, ARTM это чисто оптимизационный подход, но байесовские модели могут быть к нему адаптированы. При этом значительно упрощаются математические выкладки, появляется возможность собирать композитные модели с требуемыми свойствами из готовых модулей — регуляризаторов.

Обзор охватывает семь основных направлений, по которым развивалась теория и практические применения ARTM: 1) комбинирование регуляризаторов и выработка проблемно-ориентированных стратегий регуляризации в задачах информационного поиска и цифровых гуманитарных исследованиях, 2) реализация модульной библиотеки **BigARTM** для параллельной обработки больших текстовых коллекций, 3) построение тематических иерархий, 4) обработка мультимодальных данных, включающих n -граммы, языки, авторов, время, ссылки и т. д., 5) обработка транзакционных данных, описывающих сложные непарные взаимодействия объектов различной природы, в частности, банковские транзакции, 6) обработка текста как последовательности термов за пределами гипотезы «мешка слов», 7) автоматизация подбора коэффициентов регуляризации и других гиперпараметров. Обозначается место и роль вероятностного тематического моделирования в эпоху больших языковых моделей. Обсуждаются открытые проблемы и направления будущих исследований.

1 Введение

Тематическое моделирование — это одна из технологий *обработки естественного языка* (natural language processing, NLP). Она появилась в конце 90-х годов [43, 13] и до настоящего времени находит многочисленные применения и продолжает развиваться.

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

Тематическое моделирование относится к *машинному обучению без учителя* (unsupervised learning) и к *мягкой кластеризации текстов* (soft clustering of text). 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. Они являются основными параметрами PTM. Обучить PTM означает оценить эти параметры по входной коллекции текстовых документов.

Как и любая вероятностная языковая модель, PTM предсказывает появление слов в документах. Однако в этой роли тематические модели не настолько хороши, как современные *большие языковые модели* (large language model, LLM) на основе глубоких нейронных сетей, такие как BERT [28], GPT [17] и многие последующие. Тематические модели не способны генерировать качественный связный текст, зато кластеризацию больших текстовых коллекций они выполняют лучше и быстрее, не требуя мощных и дорогих вычислителей.

Тематическая модель, как и нейросетевая, преобразует слова и тексты в их *векторные представления* или *эмбединги* (embedding). Нейросетевые эмбединги не интерпретируемы, мы не понимаем смысла координат в этих векторах. Тематический эмбединг — это вектор вероятностей тем. Каждая тема *интерпретируема*. Это означает, что наиболее частотные слова темы связаны друг с другом по смыслу, образуя *лексическое ядро темы*. Для такой темы возможно подобрать название и составить связное текстовое описание [20, 67].

Тематические эмбединги, как и нейросетевые, могут использоваться (и раньше часто использовались) для информационного поиска, классификации, сегментации, суммаризации документов или их фрагментов. Однако по качеству решения этих задач тематические модели заметно уступают современным нейросетевым LLM.

Наиболее ценная способность тематической модели заключается в том, что она разбивает коллекцию на заданное число интерпретируемых тем. Пользователь может быть уверен, что никакие темы не упущены, поскольку это результат кластеризации. Подобное сочетание полноты, интерпретируемости и скорости пока не достижимо для LLM, и в данной нише PTM оставляет за собой лидирующую позицию.

Интерпретируемость важна для большинства применений тематического моделирования в *разведочном информационном поиске* (exploratory search) [81, 41, 80, 52], в том числе в поиске научных, патентных, новостных, юридических документов. Вторая важная область применений — это *цифровые гуманитарные исследования* (digital humanities) в таких областях как филология, культурология, психология, социология, политология, история, журналистика, маркетинг [16, 107, 53, 79].

Тематическая модель помогает понять, «о чём все эти тексты», не читая их, а также отфильтровать ту часть коллекции, которая представляет интерес для более детального анализа. 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 массива текстов.

Байесовское обучение надолго стало основной математической техникой в тематическом моделировании после работ [13, 14]. В байесовском подходе оцениваются не сами параметры модели, а их апостериорные распределения. Это существенно усложняет вывод модели, но на практике оказывается избыточным. Для приложений в большинстве случаев нужны сами параметры, а не их распределения.

Аддитивная регуляризация тематических моделей (ARTM), впервые предложенная в [94], основана на более простом и естественном оптимизационном подходе, который повсеместно принят в машинном обучении и языковом моделировании. Бай-

есовские тематические модели переносятся в ARTM без потерь функциональности. В некоторых случаях это требует определённого переосмысления модели, но всякий раз математические выкладки упрощаются радикально [93, 96].

Регуляризатор — это дополнительный критерий оптимизации, наделяющий модель определённым свойством. Аддитивная регуляризация даёт свободу переносить регуляризаторы с одних моделей на другие, получая в результате композитные модели с требуемым набором свойств. Модульный подход ARTM поддерживается библиотекой тематического моделирования с открытым кодом **BigARTM**. Обучение композитных моделей реализуется одним общим ЕМ-алгоритмом, к которому регуляризаторы подключаются как модули.

Цель данного обзора — представить ретроспективу развития и современное состояние теории, методов и приложений ARTM, обозначить открытые проблемы вероятностного тематического моделирования и наметить пути их решения.

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) [43, 44].

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 $(w_1, \dots, w_{n_d})$ из словаря W . Обычно в роли термов выступают слова, но также это могут быть словосочетания или части слов, в зависимости от токенизации, применённой на стадии предварительной обработки текста. Термами могут быть также нетекстовые токены, такие как авторы, ссылки, источники, время и т. д.

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; порядок термов в тексте игнорируется.

Языковая 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

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

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

$$\sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \varphi_{wt} \theta_{td} \rightarrow \max_{\Phi, \Theta}, \quad (2)$$

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

$$\sum_{w \in W} \varphi_{wt} = 1, \quad \varphi_{wt} \geq 0; \quad \sum_{t \in T} \theta_{td} = 1, \quad \theta_{td} \geq 0. \quad (3)$$

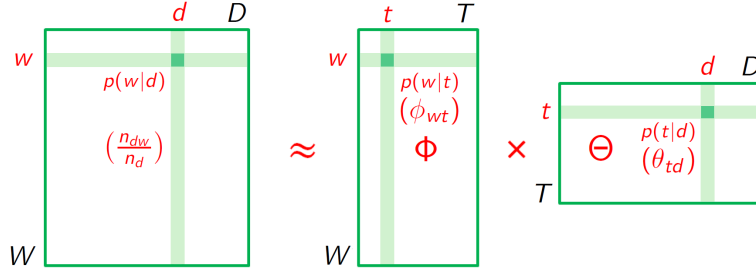


Рис. 1: Построение тематической модели текстовой коллекции сводится к вычислению низкорангового стохастического матричного разложения. Матрица частот термов в документах $p(w|d)$ приближается произведением матрицы Φ вероятностей термов в темах $p(w|t)$ и матрицы Θ вероятностей тем в документах $p(t|d)$. Все три матрицы стохастические, т.е. имеют неотрицательные и нормированные столбцы.

Оптимизационная задача (2)–(3) представляет собой задачу низкорангового стохастического матричного разложения, рис. 1. 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 Ён и Майкл Джордан в [13, 14] 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.

Метод *максимума апостериорной вероятности* (maximum a posteriori probability, MAP) для модели LDA приводит к оптимизационной постановке, в которой к основному критерию логарифма правдоподобия (2) добавляются критерии регуляризации:

$$\begin{aligned} \sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \varphi_{wt} \theta_{td} + \sum_{t \in T} \sum_{w \in W} (\beta_w - 1) \ln \varphi_{wt} + \\ + \sum_{d \in D} \sum_{t \in T} (\alpha_t - 1) \ln \theta_{td} \rightarrow \max_{\Phi, \Theta}, \end{aligned} \quad (4)$$

где $\alpha_t > 0$, $\beta_w > 0$ — *гиперпараметры* априорных распределений Дирихле. На практике их обычно задают вручную, хотя известны способы их оптимизации [106].

При $\beta_w = 1$, $\alpha_t = 1$ распределение Дирихле совпадает с равномерным распределением, а модель LDA переходит в PLSA [39].

При $\beta_w > 1$, $\alpha_t > 1$ регуляризаторы имеют сглаживающий эффект, приближая распределения φ_t и θ_d к заданным распределениям β и α соответственно.

При $0 < \beta_w < 1$, $0 < \alpha_t < 1$ распределения φ_t и θ_d , наоборот, отдаляются от распределений β и α . Регуляризатор имеет разреживающий эффект и может приводить к обнулению малых условных вероятностей в матрицах параметров Φ и Θ . К сожалению, требование строгой положительности β_w и α_t в распределениях Дирихле ограничивает возможности управления разреженностью в матрицах Φ и Θ .

Модель LDA до сих пор остаётся наиболее цитируемой в тематическом моделировании. Многие последующие модели были развитием и специализацией LDA для конкретных ситуаций в рамках байесовского обучения, см. обзоры [26, 12, 53, 23].

Хотя LDA давно морально устарела, она остаётся наиболее частым выбором для практиков. Этот парадокс объясняется, видимо, тем, что байесовское обучение так

и не смогло выработать гибкого фреймворка, в котором тематические модели было бы легко заменять, оценивать, выбирать и комбинировать.

3 От байесовского обучения к аддитивной регуляризации

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

Байесовский вывод оценивает апостериорное распределение параметров $p(\Omega | X, \gamma)$ по заданному априорному распределению $p(\Omega | \gamma)$ по формуле Байеса:

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

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:

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

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

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

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

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

где τ_i — коэффициенты регуляризации, отвечающие за балансировку критериев. На практике их подбирают экспериментальным путём.

Несмотря на различия между байесовским обучением и аддитивной регуляризацией, получаемые в результате ЕМ-подобные алгоритмы отличаются незначительно. Впервые этот факт был замечен, по всей видимости, в [9] 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 [100, 77, 101]. Правдоподобие модели LDA действительно выше, но это достигается благодаря сглаживанию оценок вероятности для редких слов и редких тем, которые наименее важны и не влияют на интерпретируемость модели. Эти исследования показали, что LDA является довольно слабым регуляризатором. Регуляризация может и должна быть усилена более содержательными лингвистическими и прикладными требованиями.

Аддитивная регуляризация тематических моделей (additive regularization for topic modeling, ARTM) была предложена Константином Воронцовым в [94]. Первоначально рассматривалась регуляризация модели PLSA (2)–(3):

$$\sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \varphi_{wt} \theta_{td} + R(\Phi, \Theta) \rightarrow \max_{\Phi, \Theta}, \quad R(\Phi, \Theta) = \sum_i \tau_i R_i(\Phi, \Theta). \quad (5)$$

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 «основной леммы» о максимизации гладкой функции на единичных симплексах [96].

Введём операцию norm , которая преобразует произвольный вектор $(x_i)_{i \in I} \in \mathbb{R}^{|I|}$ в неотрицательный нормированный вектор $(p_i)_{i \in I}$:

$$p_i = \text{norm}_{i \in I}(x_i) = \frac{\max(0, x_i)}{\sum_{j \in I} \max(0, x_j)}, \quad \text{для всех } i \in I.$$

Если $x_i \leq 0$ для всех $i \in I$, то результат norm определим как нулевой вектор.

Лемма 1. [96]. 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

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

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

$$\omega_{ij} = \text{norm}_{i \in I_j} \left(\omega_{ij} \frac{\partial f}{\partial \omega_{ij}} \right). \quad (6)$$

Для решения системы уравнений удобно использовать метод простой итерации. Сначала задаются начальные приближения для векторов ω_j , затем их значения многократно обновляются по формуле (6). Итерационный процесс отличается лишь тремя нюансами от градиентного метода безусловной максимизации

$$\omega_{ij} = \omega_{ij} + \eta \frac{\partial f}{\partial \omega_{ij}}.$$

Во-первых, мультипликативный градиентный шаг используется вместо аддитивного. Во-вторых, отсутствует параметр градиентного шага η . В-третьих, операция norm проецирует полученный вектор на единичный симплекс.

Теорема 1. [104]. Пусть функция $R(\Phi, \Theta)$ непрерывно дифференцируема. Тогда локальным экстремумом задачи (5) с ограничениями (3) являются матрицы (Φ, Θ) , ненулевые столбцы которых удовлетворяют следующей системе уравнений со вспомогательными переменными p_{tdw} , где $d \in D$, $w \in W$, $t \in T$:

$$p_{tdw} = \text{norm}_{t \in T}(\varphi_{wt} \theta_{td}); \quad (7)$$

$$\varphi_{wt} = \text{norm}_{w \in W} \left(n_{wt} + \varphi_{wt} \frac{\partial R}{\partial \varphi_{wt}} \right); \quad n_{wt} = \sum_{d \in D} n_{dw} p_{tdw}; \quad (8)$$

$$\theta_{td} = \text{norm}_{t \in T} \left(n_{td} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right); \quad n_{td} = \sum_{w \in W} n_{dw} p_{tdw}. \quad (9)$$

Доказательство легко следует из леммы 1.

Систему уравнений (7)–(9) удобно решать методом простых итераций, который принято называть ЕМ-алгоритмом (expectation–maximization). Сначала выбираются начальные приближения $\varphi_{wt}, \theta_{td}$, затем в цикле до сходимости чередуются *E-step* (7) и *M-step* (8)–(9).

Интерпретации вспомогательных переменных:

$p_{tdw} = p(t | d, w)$ оценивает вероятность темы t для термина w в документе d ;

n_{wt} оценивает, сколько раз терм w относится к теме t во всей коллекции;

n_{td} оценивает, сколько раз термины документа 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 с помощью регуляризации является полезной инструментальной особенностью ARTM.

В последующих работах [102, 103, 104] байесовские тематические модели были массово переформулированы в виде регуляризаторов $R(\Phi, \Theta) = \ln p(\Phi, \Theta | \gamma)$: сглаживание и разреживание (обобщение модели LDA), частичное обучение, декоррелирование тем для повышения их различности, классификация документов и другие [95, 59, 96].

В частности, модель LDA была переформулирована без введения априорных распределений Дирихле, как регуляризатор кросс-энтропии для сглаживания или разреживания столбцов матриц Φ и Θ . Тем самым было снято ограничение на сильное разреживание в (4), которое возникало в LDA из-за строгой положительности гиперпараметров $\alpha_t > 0$ и $\beta_w > 0$ в распределении Дирихле [104].

Модель LDA играет особую роль в байесовском тематическом моделировании благодаря сопряжённости распределения Дирихле с мультиномиальным распределением. Свойство сопряжённости сильно упрощает байесовский вывод. Модель LDA не имеет убедительных лингвистических обоснований, но благодаря математическому удобству она стала общепринятой теоретической базой вероятностного 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

Stages	Bayesian Inference for PTMs	ARTM	
<i>Requirements analysis:</i>	Requirements analysis	Requirements analysis	
<i>Model formalization:</i>	Generative model design	predefined criteria	user-defined criteria
<i>Model inference:</i>	Bayesian inference for the generative model (VI, GS, EP)	One regularized EM-algorithm for any combination of criteria	
<i>Model implementation:</i>	Researchers coding (Matlab, Python, R)	Production code (C++)	
<i>Model evaluation:</i>	Researchers coding (Matlab, Python, R)	predefined measures	user-defined measures
<i>Deployment:</i>	Deployment	Deployment	

conventions: ::: not unified stages ::: ::: unified stages :::

Рис. 2: ARTM позволяет унифицировать основные этапы тематического моделирования, в отличие от байесовского обучения. Обработка и хранение данных, быстрый параллельный ЕМ-алгоритм, модульные библиотеки регуляризаторов и метрик качества имеют общую реализацию для широкого класса моделей и их композиций.

и тем же итерационным алгоритмом, который реализуется и отлаживается один раз, а регуляризаторы добавляются в него как отдельные вычислительные модули.

Байесовский подход не поддерживает сопоставимой гибкости и модульности. Наоборот, байесовский вывод уникален для каждого априорного распределения, что приводит всякий раз к новому алгоритму, который необходимо заново кодировать и отлаживать, рис. 2.

Илья Ирхин в теоретической работе [51] 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 надо подставить в правые части формул М-шага вместо φ_{wt} и θ_{td} , взятых с предыдущей итерации, их несмещённые (соответствующие максимуму правдоподобия при $R = 0$) оценки φ_{wt}^* и θ_{td}^* , взятые с текущей итерации:

$$\varphi_{wt} = \text{norm}_{w \in W} \left(n_{wt} + \varphi_{wt}^* \frac{\partial R(\Phi^*, \Theta^*)}{\partial \varphi_{wt}} \right); \quad \varphi_{wt}^* = \text{norm}_{w \in W} (n_{wt}); \quad (10)$$

$$\theta_{td} = \text{norm}_{t \in T} \left(n_{td} + \theta_{td}^* \frac{\partial R(\Phi^*, \Theta^*)}{\partial \theta_{td}} \right); \quad \theta_{td}^* = \text{norm}_{t \in T} (n_{td}). \quad (11)$$

Эксперименты в [51] показали, что данная модификация повышает значение регуляризованного правдоподобия, причём итерационный процесс гораздо быстрее (как правило, уже со второй итерации) входит в режим монотонного увеличения регуляризованного правдоподобия. Чем больше коэффициент регуляризации, тем сильнее проявляется эффект улучшения сходимости.

Анна Потапенко в серии экспериментов [102, 103, 104] показала, что композиция шести регуляризаторов (сглаживание, разреживание, декоррелирование и отбор тем) улучшает одновременно несколько метрик качества (когерентность, разреженность, чистота и контрастность тем) ценой незначительного ухудшения правдоподобия. Регуляризаторы могут вводиться в композицию как вместе, так и по очереди.

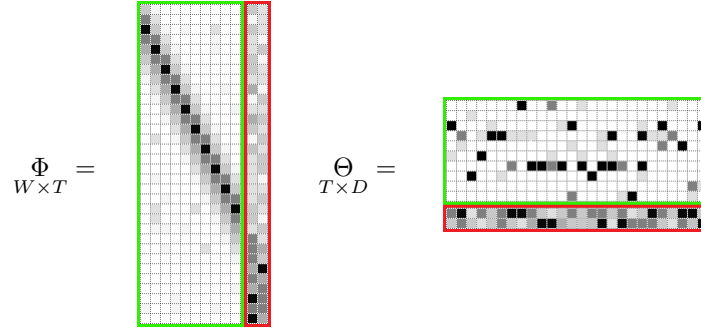


Рис. 3: Структура разреженности матриц Φ и Θ с предметными и фоновыми темами.

Управляемое изменение коэффициентов регуляризации τ_i в ходе итерационного процесса называется *стратегией регуляризации*. Некоторые эмпирические рекомендации по стратегии регуляризации были выработаны в [104]. Декоррелирование и сглаживание лучше включать с первой же итерации. Разреживание лучше начинать 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 неоднократно использовались в последующих работах [105, 5, 47, 46, 84].

Декоррелирование впервые было предложено в модели TWC-LDA (topic-weak-correlated LDA) в рамках байесовского подхода [88]. В ARTM ему соответствует критерий регуляризации

$$R(\Phi) = - \sum_{t \in T} \sum_{s \in T \setminus t} \sum_{w \in W} \varphi_{wt} \varphi_{ws}.$$

Соответствующее априорное распределение не является сопряжённым к мультиномиальному, поэтому байесовский вывод сталкивается с техническими трудностями. В ARTM аддитивная поправка к формуле М-шага выводится в одну строку.

Регуляризатор декоррелирования показал себя как один из наиболее полезных во многих приложениях. Он усиливает различность тем как столбцов матрицы Φ , препятствует образованию тем-дубликатов, способствует разреживанию матрицы Φ и группированию общей лексики в отдельные темы. Последнюю тенденцию можно усилить, если разделить множество тем на предметные и фоновые в явном виде, $T = S \sqcup B$, рис. 3.

Фоновые темы $t \in B$ собирают в себе общую лексику языка. Сглаживание применяется к соответствующим столбцам матрицы Φ и строкам матрицы Θ . Тематическая модель с несколькими фоновыми темами обобщает робастные модели с единственным фоновым распределением [21, 77].

Предметные темы $t \in S$ содержат термины предметных областей. Разреживание применяется к соответствующим строкам матрицы Θ . Декоррелирование можно применять ко всем темам, при этом разреживать предметные темы как столбцы матрицы Φ уже не обязательно, для этого достаточно декоррелирования.

Завершая данный раздел, отметим, что не существует теоретических или практических препятствий к тому, чтобы переписывать в терминах ARTM любые байесовские тематические модели, получая возможность их комбинирования практически

в любых сочетаниях. Кроме того, переход от байесовского вывода к градиентной оптимизации в рамках ARTM имеет большой потенциал для симбиоза вероятностных тематических моделей с нейросетевыми моделями языка. Ранние попытки такого симбиоза наталкивались на значительные технические трудности [111].

4 От PLSA к проблемно-ориентированной регуляризации

В данном разделе сведены исследования, в которых ARTM применялся для тематического анализа текстов из разных предметных областей. Иногда это требовало введения нового регуляризатора, как в [83]. Однако в большинстве случаев целям анализа и требованиям заказчика удовлетворял выбор из готовых регуляризаторов, к тому времени уже реализованных в библиотеке BigARTM. Каждый прикладной кейс подтверждал, что использование регуляризаторов улучшает проблемно-ориентированные метрики качества по сравнению с простыми моделями PLSA и LDA. Некоторые работы показали необходимость тщательного дизайна экспериментов по подбору коэффициентов регуляризации, числа тем и других гиперпараметров. Эти наблюдения легли в основу более поздних разработок TopicNet и AutoTM.

Преимущества ARTM проявляются во многих приложениях, где тематическая модель используется как вспомогательный инструмент анализа текстов. Тематические векторы $p(t | \cdot)$ часто используются в качестве признаков описаний текстовых объектов в задачах информационного поиска и предсказательного моделирования — классификации, регрессии, ранжирования.

Необходимость предсказывать числовую величину по текстовому документу возникает во многих приложениях [66]: 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 биржевого курса актива по экономическим новостям и т. д.

Евгений Соколов и Лев Боголюбский в [83] предложили использовать критерий наименьших квадратов, обычный для регрессионных моделей, в качестве регуляризатора для тематической модели ARTM. Фактически, в итерационном EM-алгоритме две модели поочерёдно приспособляются друг к другу. Тематическая модель формирует распределения $p(t | d)$, которые используются в качестве признаков векторов документов d в линейной регрессии. Регрессионная модель воздействует на темы через регуляризатор таким образом, чтобы тематические признаки оказывались как можно более информативными именно для данной регрессионной задачи.

Сергей Николенко в [74] 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 и др. в [5, 6] применили ARTM для выявления тематики национальностей и межнациональных отношений в постах социальных сетей. Для поиска этно-релевантных 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 [15, 60] подход ARTM помогал социологам находить больше интерпретируемых тем, в том числе полиэтнические темы, выявляющие особенности межэтнических отношений.

В [65] модель ARTM строилась по 13.6М репозиториям программного обеспечения на GitHub с целью понимания и категоризации тематики проектов с открытым кодом. Были построены 256 тем, которые удалось распределить вручную по 6 крупным категориям. Для повышения интерпретируемости тем использовались регуляризаторы разреживания.

В серии работ [63, 61, 64, 62] модель ARTM использовалась для выявления тематической кластерной структуры коллекции научных статей из библиотеки OnePetro, относящихся к разработке нефте-газовых месторождений. Сначала число тем фиксировалось по числу рубрик библиотечного классификатора, взятых для формирования исходной коллекции. С помощью регуляризаторов выделялись две фоновые темы. В [62] была предложена модифицированная метрика качества кластеризации cDBI (Davies–Bouldin Index), по которой удалось определить оптимальное число тем. В [64] тематическая модель использовалась для выявления способа перевода текстов с русского языка на английский на основе различий в переводах терминологии.

В [78] ARTM применялся для выявления клинических маршрутов на основе журнала событий в больнице. Траектория пациента как последовательность стандартных действий (регистрация, диагностические и лечебные мероприятия, выписка) аналогична последовательности слов в документе. Темы соответствуют клиническим схемам или маршрутам. Тематическая модель с регуляризаторами декоррелирования и разреживания позволила выявить и интерпретировать больше маршрутов, чем жесткая кластеризация временных рядов. Тематическое моделирование рассматривается в данной работе как инструмент для разработки персонализированных подходов, определения групп риска и рекомендаций профилактических мер на основе данных гетерогенного потока пациентов в городских медицинских учреждениях.

В [109, 108, 70] модель ARTM с регуляризаторами декоррелирования и разреживания строилась по миллиону новостных сообщений за 2020 год. Для выделения тем, связанных с медициной, пандемией и здравоохранением, был предложен каскадный метод, состоящий в следующем. Последовательно строились четыре модели. В каждой модели эксперты отбирали релевантные темы, и документы, содержащие хотя бы 5% релевантной темы, отбирались для построения следующей модели. На выходе каскада осталось менее 12% документов, по которым были построены релевантные темы и исследована их динамика во времени.

В [69] модель ARTM использовалась для анализа программы импортозамещения на основе патентных данных (152К документов). Программа включает планы для 22 отраслей и содержит более 1500 продуктов и технологий для предлагаемого импортозамещения. Для фиксации релевантных тем, соответствующих отраслям экономики, были введены 22 регуляризатора сглаживания. Релевантные темы сглаживались по словарям, составленным из терминов отраслевых планов импортозамещения. Биграммы выделялись в отдельную модальность. Стратегия массового патентного поиска на основе тематического моделирования позволила получить не только всестороннюю картину эффективности программы в целом, но и более подробную информацию о том, какие группы продуктов и технологий были запатентованы.

В [84] модель ARTM строилась для оценивания уровней сложности тематической структуры учебных текстов. Регуляризаторы декоррелирования и разреживания использовались для повышения интерпретируемости тем.

Многие приложения тематического моделирования связаны с обработкой динамически растущих текстовых коллекций, в которых каждый документ имеет метку времени. К такому типу данных относятся социальные медиа, новостные потоки, научные архивы, патентные базы.

Николай Герасименко и др. в [37, 36, 38] 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 [32, 25, 33] 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 [87] 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.

5 От теории ARTM к реализации BigARTM

Проект BigARTM по созданию библиотеки тематического моделирования с открытым кодом на основе теории ARTM был анонсирован в 2014 году [99, 98]. BigARTM реализует многопоточный пакетный EM-алгоритм для эффективной обработки больших текстовых коллекций. Имеется две встроенных расширяемых библиотеки: одна с регуляризаторами, вторая с обработчиками (scores) для оценивания качества модели и вывода промежуточных результатов моделирования.

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 сходится даже когда не все пакеты обработаны по одному разу. В таких случаях мы говорим про *онлайн-овый* 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 *оффлайнowymi*. BigARTM предоставляет оба варианта реализации, онлайнный и оффлайнный. 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 [35]. 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 [59] 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 [8]. 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 [7] 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.

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 [105] исследовал возможность применения энтропийного регуляризатора разреживания, ранее предложенного в [103], для удаления незначимых тем. В байесовском подходе для определения числа тем используют непараметрическую модель иерархического процесса Дирихле (hierarchical Dirichlet process, HDP) [89]. В обеих моделях имеется параметр, от которого число тем зависит существенно: коэффициент регуляризации τ в ARTM и коэффициент концентрации γ в HDP.

Эксперименты в [105] 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 устой-

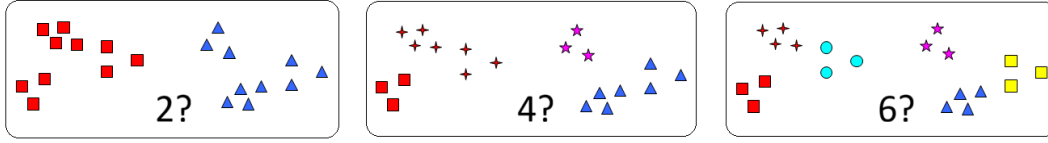


Рис. 4: Выбор числа тем в тематических моделях столь же неоднозначен, как выбор числа кластеров в задачах кластеризации.

чивее. 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 [105] 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 различных подходов к определению числа тем, включая аддитивную регуляризацию, было проведено в [18]. Оно подтвердило, что число тем скорее зависит от модели и метода оценивания, чем при-суще самой текстовой коллекции. Аналогичная проблема неоднозначности выбора числа кластеров возникает и в задачах кластеризации, рис. 4.

На практике во многих случаях гораздо продуктивнее строить тематические иерархии, в которых темы разделяются на подтемы пока это целесообразно, чем пытаться искать оптимальное число тем. При этом уровень детализации тем можно выбирать исходя из оценок качества решения целевой прикладной задачи, например поиска или классификации текстов. Падение качества может происходить как при недостаточном числе тем — из-за недостатка размерности и недообучения модели, так и при избыточном числе тем — из-за недостатка данных и переобучения модели.

Стратегии построения тематических иерархий весьма разнообразны [110]. 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 [22] 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 $p(s | t)$ — условные вероятности подтем s в темах t . При построении дочернего уровня родительские темы $p(w | t)$ рассматриваются как псевдодокументы, которые просто добавляются к коллекции. Искомые распределения $p(s | t)$ считываются из столбцов матрицы Θ , соответствующих псевдодокументам, как показано на рис. 5.

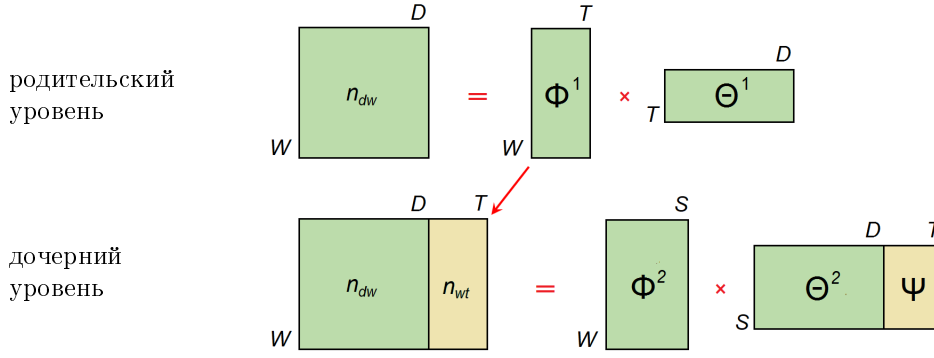


Рис. 5: Для перехода от родительского уровня с множеством тем T к дочернему уровню с подтемами S в коллекцию добавляются $|T|$ псевдодокументов с частотами термов n_{wt} . Результирующая матрица связей тем с подтемами $\Psi = (p(s | t))$ размера $S \times T$ считывается в столбцах матрицы Θ , соответствующих псевдодокументам.

Позже этот подход был улучшен в [11], реализован в BigARTM и использован во многих исследованиях [90, 48, 49, 75].

7 От слов к модальностям

Мультимодальная тематическая модель описывает документы, содержащие метаданные или нетекстовые данные 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 данных образует отдельную *модальность* со своим словарём. Слова, словосочетания, n -граммы, теги, названия, аббревиатуры — это примеры текстовых модальностей. В мультязычных 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 [99, 98, 93]. Каждый документ рассматривается как контейнер, содержащий термы различных модальностей, включая обычные слова, см. рис. 6.

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

$$\begin{aligned} \sum_{m \in M} \tau_m \sum_{d \in D} \sum_{w \in W_m} n_{dw} \ln \sum_{t \in T} \varphi_{wt} \theta_{td} + R(\Phi, \Theta) &\rightarrow \max_{\Phi, \Theta}; \\ \sum_{w \in W_m} \varphi_{wt} = 1; \quad \varphi_{wt} \geq 0; \quad \sum_{t \in T} \theta_{td} = 1; \quad \theta_{td} \geq 0; \end{aligned} \quad (12)$$

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

ЕМ-алгоритм для задачи (12) без труда выводится из леммы 1; более ранний его вывод из условий Каруша–Куна–Таккера можно найти в [93].

Анастасия Янина применила мультимодальную модель ARTM для тематического разведочного поиска в [47] и позже улучшила её в серии работ [46, 48, 49]. Качество

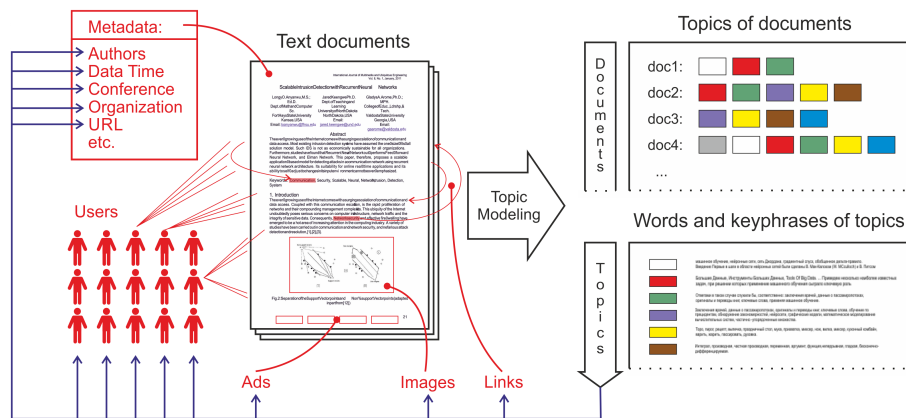


Рис. 6: Обычная тематическая модель определяет распределение тем в каждом документе $p(t|d)$ и распределение термов в каждой теме $p(w|t)$. Мультимодальная модель определяет распределения $p(w|t)$ также для термов $w \in W^m$ всех остальных модальностей $m \in M$, в том числе нетекстовых.

тематического поиска измерялось критериями точности и полноты на основе оценок ассессоров. 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 ассессорами, затем системой поиска документов по близости их тематических векторов $p(t|d)$, затем те же ассессоры 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 механизмов 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 ассессоры, 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 [46]. 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 $p(t|d)$ чётко определились релевантные темы документов. Третий этап из 8 итераций добавлял сглаживание столбцов матрицы Φ , чтобы предотвратить вырождение тем как распределений $p(w|t)$. 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, улучшала также и качество тематического поиска, хотя никакие критерии качества поиска явным образом не оптимизировались.

Применение иерархических тематических моделей в [48] 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 $p(t|d)$. 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 [48] 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.

Сравнение качества поиска на тех же данных в [49] показало, что иерархические регуляризованные мультимодальные модели ARTM достигают полноты и точности более 90%, сопоставимой с нейросетевыми эмбедингами BERT. При этом ARTM требует тщательного подбора стратегии регуляризации, тогда как BERT работает практически «из коробки». С другой стороны, модели ARTM обходятся без предобучения и намного менее требовательны к вычислительным ресурсам. Обычные тематические модели типа LDA вообще неконкурентоспособны в данной задаче.

Даниил Фельдман в [34] 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 (субъект, предикат, объект). Абляционные эксперименты показали, что каждая из трёх модальностей важна для повышения качества выделения мнений.

Аналогичная двухуровневая иерархия использовалась в [75] для маршрутизации обращений в контактный центр. Верхний уровень строился по лексике, чтобы выделить тематику обращения. Второй уровень строился по морфологическим и синтаксическим модальностям, чтобы выделить в каждой теме клиентские интенции, например, различая фразы «заблокируйте мне счёт» и «у меня заблокирован счёт».

Марина Суворова (Дударенко) в [99, 98, 29] использовала модальности языков для мультязычной 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 [38] строила мультязычную тематическую модель для поиска и классификации научных текстов 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 [38] 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.

8 От модальностей к транзакционным данным

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

Тематический анализ банковских транзакционных данных [30] 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 [58] перенёс эту методику на анализ банковских транзакционных данных юридических лиц. В роли документов выступают компании, терминами 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 [58] 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 [96]. 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:

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

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 :

$$\begin{aligned} \sum_{k \in K} \tau_k \sum_{(d,x) \in D_k} n_{kdx} \ln \left(\sum_{t \in T} \theta_{td} \prod_{v \in x} \varphi_{vt} \right) + R(\Phi, \Theta) \rightarrow \max_{\Phi, \Theta}; \\ \sum_{v \in W_m} \varphi_{vt} = 1, \varphi_{vt} \geq 0; \quad \sum_{t \in T} \theta_{td} = 1, \theta_{td} \geq 0; \end{aligned} \quad (13)$$

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 постановка задачи является, по всей видимости, одной из самых общих в тематическом моделировании [96]. EM-алгоритм для задачи (13) без труда выводится из леммы 1. Этот алгоритм реализован в библиотеке **BigARTM**.

В [96, 97] предложены транзакционные модели для рекомендательных систем. Здесь клиенты или пользователи (users) выступают в роли документов, объекты или товары (items) выступают в роли слов. Основной тип парных транзакций «клиент u предпочёл товар i » является непосредственным аналогом «документ u содержит слово i ». Тематическое моделирование было введено Т. Хофманном практически одновременно для анализа текстов [43] и для рекомендаций [45].

Рекомендательные системы часто испытывают проблему холодного старта, когда новому клиенту нечего порекомендовать, поскольку мы не имеем истории его предпочтений, а также когда новый товар некому порекомендовать, поскольку его ещё никто не выбирал. Для решения этой проблемы вводятся ещё две модальности и данные о парных транзакциях ещё двух типов: «объект i имеет свойство f » и «клиент u имеет атрибут a ». Наконец, введение данных о времени, дилере, целях клиента порождает n -арные транзакции.

Когда объём данных увеличивается динамически, ни одна из модальностей не может служить в качестве контейнера. Эта ситуация структурно отличается от динамических текстовых коллекций, в которых может увеличиваться число документов, однако каждый документ навсегда остаётся в первоначальном виде. Трудно представить ситуацию, когда слова добавляются в документ также динамически. В рекомендательных системах данные о клиенте, как и данные об объекте, могут со временем пополняться. Для этой ситуации в [96, 97] предложены симметризованные транзакционные тематические модели на основе ARTM.

9 От мешка слов к тематическому вниманию

Гипотеза «мешка слов» является одним из самых критикуемых предположений тематического моделирования. В рамках ARTM были разработаны шесть 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. Например, триграмма «метод опорных векторов» явно указывает на тему «машинное обучение», в то время как эти же три слова по-отдельности «опорный», «вектор», «метод» имеют намного более размытую тематику.

В ряде работ [48, 38, 37, 36] для построения словаря n -грамм использовался быстрый алгоритм TopMine [31]. Он выделяет статистически значимые коллокации за время, линейное по длине текстовой коллекции. TopMine применялся на стадии формирования входных данных для BigARTM.

2. *Тематическая модель дистантной сочетаемости* основана на предположении, что если два термина 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 [76] предложили формировать для каждого термина 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 [68] и существенно превосходит обычные тематические модели. Это означает, что для тематических векторов, как и для эмбедингов word2vec, выполняются тождества между ассоциативными парами слов, например,

$$\begin{aligned} p(t \mid \text{король}) - p(t \mid \text{королева}) &= p(t \mid \text{мужчина}) - p(t \mid \text{женщина}); \\ p(t \mid \text{Москва}) - p(t \mid \text{Пекин}) &= p(t \mid \text{Россия}) - p(t \mid \text{Китай}). \end{aligned}$$

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 [24], специально разработанную для поиска семантически близких документов.

Также в [76] 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 термина была определена его окрестность.

В [90] 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

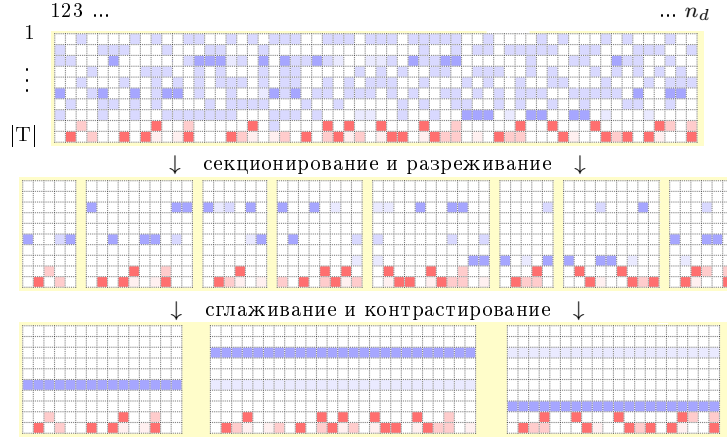


Рис. 7: Эвристическая пост-обработка матрицы $\Pi = (p(t | d, w_i))_{T \times n}$ на Е-шаге. Сначала документ секционируется по предложениям, в каждом предложении определяется основная предметная тема. Затем соседние предложения с одинаковой основной темой объединяются в сегменты s , для которых определяются распределения $p(t | d, s)$. Фоновые темы (две нижние строки) не участвуют в пост-обработке.

слова предложения порождаются одной определённой темой, не раз эксплуатировалась в литературе [112, 10]. Транзакционная тематическая модель позволяет связывать не только слова предложений, но и любые подмножества слов x , предположительно генерируемых одной общей темой из распределения $p(t | d)$. 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 Е-шага.* Согласно (7), ЕМ-алгоритм на Е-шаге вычисляет тематические векторы $p(t | d, w)$ через параметры модели Φ, Θ . Последовательность тематических векторов образует $|T| \times n$ -матрицу $\Pi(\Phi, \Theta)$, где n — суммарная длина всех документов коллекции. Столбцы в матрице Π идут в том же порядке, что и термины в исходных документах.

В [95] были введены регуляризаторы вида $R(\Pi(\Phi, \Theta))$, чтобы воздействовать на матрицу Π непосредственно. Было показано, что воздействие такого регуляризатора на модель сводится к преобразованию векторов $p(t | d, w)$ перед их подстановкой в формулы М-шага. Отсюда название — пост-обработка Е-шага.

Николай Скачков в [82] 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 эвристические преобразования, например, секционирование, контрастирование, разреживание, сглаживание и сегментация, как показано на рис. 7.

Позже было доказано [97], что любое преобразование матрицы Π , выполненное после Е-шага и перед М-шагом, эквивалентно некоторому регуляризатору, который не обязательно формулировать явно в виде критерия $R(\Pi)$. Пост-обработка Е-шага оставляет большую свободу выбора чисто алгоритмических эвристик для учёта естественного порядка слов в тексте.

5. *Быстрая линейная векторизация документов* была предложена в [50], чтобы исключить матрицу Θ из тематической модели, избежать переобучения на коротких документах и ускорить вычисления. Было предложено определять тематический вектор документа как среднее арифметическое тематических векторов его термов:

$$\theta_{td}(\Phi) \equiv p(t | d) = \frac{1}{|I_d|} \sum_{i \in I_d} p(t | w_i), \quad (14)$$

где I_d — множество позиций, занимаемых термами документа d в сквозной нумерации термов $\{w_1, \dots, w_n\}$ коллекции. Равенство (14) связывает параметры модели зависимостью $\Theta = \Theta(\Phi)$, фактически играя роль регуляризатора. Эксперименты на трёх текстовых коллекциях подтвердили, что такая регуляризация одновременно улучшает когерентность, различность и разреженность тем [50].

Линейная векторизация может быть применена как к целому документу, так и к любому его фрагменту: предложению, абзацу, секции, окрестности термина. При таком применении уже нельзя говорить, что модель игнорирует естественный порядок слов, хотя суммирование (14) по-прежнему является «мешком слов». Именно от этой модели проще всего сделать шаг в сторону тематических моделей внимания.

6. *Локальным контекстом* термина w_i называется выбранная из документа подпоследовательность окружающих его термов $C_i \subset \{1, \dots, n\}$. 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. Контекст не может выходить за пределы документа, захватывая термы других документов.

Тематическая модель локального контекста предсказывает появление термина в i -й позиции по его контексту:

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

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

Тематическая модель внимания получается из модели локального контекста, если определить тематический вектор контекста $p(w | C_i)$ по аналогии с (14) как среднее взвешенное тематических векторов его термов:

$$p(t | C_i) = \sum_{c \in C_i} \alpha_{ci} p(t | w_c); \quad \sum_{c \in C_i} \alpha_{ci} = 1; \quad \alpha_{ci} \geq 0; \quad (15)$$

где α_{ci} — *коэффициенты внимания*, показывающие, насколько важен (с каким весом учитывается) терм w_c из контекста C_i для термина w_i в позиции i . После подстановки тематическая модель локального контекста принимает вид функции только от тематических векторов $\varphi_{tw} = p(t | w)$, матрица Θ исключается из модели:

$$p(w | C_i) = \sum_{t \in T} \varphi_{tw} \frac{p(w)}{p(t)} \sum_{c \in C_i} \alpha_{ci} \varphi_{twc}. \quad (16)$$

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

$$\sum_{i=1}^n \ln p(w_i | C_i) + R(\Phi) \rightarrow \max_{\Phi} \quad (17)$$

при ограничениях неотрицательности и нормировки:

$$\sum_{t \in T} \varphi_{tw} = 1, \quad \varphi_{tw} \geq 0. \quad (18)$$

Лемма 1, применённая к этой оптимизационной задаче, даёт систему уравнений, решение которой методом простой итерации приводит к ЕМ-алгоритму. Опуская доказательство и эвристические упрощения, слабо влияющие на качество модели, запишем ЕМ-алгоритм:

$$\begin{aligned} p_{ti} &= \text{norm}_{t \in T} \left(\frac{\varphi_{tw_i}}{p(t)} \sum_{c \in C_i} \alpha_{ci} \varphi_{tw_c} \right); \quad p(t) = \sum_{w \in W} \varphi_{tw} p(w); \\ \varphi_{tw} &= \text{norm}_{t \in T} \left(n_{tw} + \varphi_{tw} \frac{\partial R}{\partial \varphi_{tw}} \right); \quad n_{tw} = \sum_{i=1}^n p_{ti} [w_i = w]. \end{aligned}$$

Отметим, что семейство моделей локального контекста оставляет свободу определения контекста C_i и коэффициентов внимания α_{ci} .

Вычисление контекстно-зависимого тематического вектора $p_{ti} = p(t | w_i, C_i)$ по контекстно-независимому вектору $\varphi_{tw_i} = p(t | w_i)$ того же термина w_i во многом аналогично известным нейросетевым моделям — свёрточной языковой модели GCNN (gated convolutional neural network) [27] и модели внимания «запрос–ключ–значение» (query–key–value) [91]. Во всех трёх случаях вычисляется взвешенная сумма векторов из контекста $p(t | w_c)$, $c \in C_i$, называемых *значениями* (value). Веса в этой сумме являются нормированные по контексту коэффициенты внимания α_{ci} , которые могут нести в себе информацию о близости, как позиционной, так и семантической, вектора термина w_c из контекста, называемого *ключом* (key) и вектора термина w_i , называемого *запросом* (query).

При этом есть и существенные отличия тематического внимания.

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 нейросетевой модели внимания векторы запросов, ключей и значений умножаются на матрицы линейных преобразований, оптимизируемые при обучении сети. Такая параметризация существенно обогащает модель, однако тематическая модель внимания в предложенном виде не имеет обучаемых параметров.

В-третьих, в предложенной модели коэффициенты внимания α_{ci} могут выражать лишь позиционную близость термов w_c и w_i , в то время как нейросетевая модель внимания вычисляет α_{ci} через семантическую близость (например, скалярное произведение) эмбедингов w_c и w_i . Отчасти это компенсируется тем, что тематические векторы ключа $p(t | w_c)$ и запроса $p(t | w_i)$ перемножаются покомпонентно. Запрос играет роль фильтра, пропуская только те темы, которые есть у термина w_i . В итоге взвешенная сумма формируется преимущественно из векторов-значений $p(t | w_c)$ тематически близких вектору-запросу $p(t | w_i)$.

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

10 От подбора гиперпараметров к метаобучению

Опыт построения стратегий регуляризации, накопленный в серии работ [102, 103, 104, 105, 93, 99, 5, 6, 22, 75, 46, 48, 49], был обобщён в библиотеке **TopicNet** [19]. 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 программирования¹.

Василий Алексеев и др. в [1] использовали двухуровневую тематическую иерархию для построения «банка тем» — устойчивого набора из наиболее удачных, хорошо интерпретируемых тем. Проблема неустойчивости тематических моделей известна давно [60], но на практике зачастую игнорируется. Неустойчивость проявляется, в частности, когда при повторных запусках ЕМ-алгоритма по одним и тем же данным, но из различных начальных приближений или при различной рандомизации, получаются разные наборы тем. Поэтому в [1] 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 [40]. Более ранние подходы к построению полного набора тем, основанные на построении выпуклой оболочки [86, 85] вычислительно более ресурсоёмкие.

Построение «банка тем», хотя и поддерживается библиотекой **TopicNet**, может оказаться слишком ресурсоёмким для отдельных практических задач. Поэтому в [1, 40] 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 [57]. Он основан на эволюционном алгоритме для построения стратегии адаптивного многоэтапного изменения гиперпараметров. В последующей работе [55] этот метод был дополнен суррогатной моделью для оценивания качества модели, что позволило примерно в 1,5 раза сократить время поиска гиперпараметров. В [54] генетический алгоритм был дополнен методом Нелдера–Мида для ускорения локального поиска в пространстве гиперпараметров. В [56] предложен фреймворк **AutoTM 2.0** для автоматической оптимизации гиперпараметров на основе оценивания тематических моделей как по когерентности, так и с использованием LLM в роли ассессора.

В этих работах 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 может оказаться слишком расточительной. Корректировать гиперпараметры можно было бы и чаще. Например, обработав очередной пакет данных и обновив модель, можно оценить её качество на следующем пакете, пусть приближённо, зато быстро и по нескольким критериям. Построение таких стратегий регуляризации пока остаётся открытым перспективным направлением исследований.

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

11 Открытые проблемы и направления исследований

В области вероятностного тематического моделирования накопилось некоторое количество проблем, не решённых или решённых лишь частично в рамках байесовского подхода. ARTM подсказывает новые пути их окончательного решения.

1. 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 устаревшую модель LDA.

Создание тематической модели внимания с поддержкой аддитивной регуляризации, мультимодальных и транзакционных данных, представляется наиболее перспективным подходом. Это потребует новых исследований для проверки гипотез о регуляризирующем воздействии внимания на устойчивость, когерентность, разреженность и другие характеристики качества тематической модели.

2. 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 [92]. Причина этого явления кроется в том, что для максимизации правдоподобия в задаче матричного разложения ЕМ-алгоритму выгодно выравнивать темы по объёму, поэтому распределение $p(t)$ имеет тенденцию к избыточному сглаживанию. Ранее предлагалось вводить распределение $p(t)$ в модель явным образом, воздействуя либо на столбцы матрицы Θ [106], либо на строки матрицы Φ [92]. Эффекты удаления мусорных и линейно зависимых тем наблюдались также при использовании энтропийного разреживания распределения $p(t)$ в [105] и регуляризатора декоррелирования.

Эвристические способы перевзвешивания тем, аналогичные применяемым в задачах классификации с несбалансированными классами, не ведут к желаемому результату, как было показано в [92]. Проблема несбалансированности в тематических моделях, по всей видимости, является более трудной, чем в задачах классификации.

Есть основания полагать, что тематические модели локального контекста лучше подходят для обработки несбалансированных коллекций, поскольку в них нет матрицы Θ , наиболее подверженной эффектам дисбаланса тем. Проверка этой гипотезы потребует новых экспериментов по выявлению и устранению эффектов дробления и слияния тем.

3. Проблема автоматического оценивания интерпретируемости тем некоторое время считалась закрытой циклом работ [71, 72, 73], 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 около 1% от её объёма. Другими словами, когерентность недостаточно репрезентативна [2].

Василий Алексеев и др. в [2] предложили оценивать *внутритекстовую когерентность*, собирая по тексту статистику тематической близости по всем биграммам. Внутритекстовая когерентность представляется наиболее адекватной величиной для автоматического оценивания интерпретируемости тем в тематических моделях внимания. Проверка этой гипотезы потребует новых калибровочных экспериментов, аналогичных [71, 72, 73].

4. Проблема эффективного *автоматического выделения терминов* (automatic term extraction) для формирования словаря тематической модели. 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, что не всегда практично. Открытой проблемой остаётся встраивание эффективных методов отбора тематических словосочетаний в алгоритмы тематического моделирования. Для пакетных и онлайн-версий ЕМ-алгоритма эта задача пока не имеет общепринятого решения.

5. Проблема краткого изложения или *суммаризации тем* (topic summarization). 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* (automatic topic labeling) была впервые поставлена и решена в [67]. Однако задача суммаризации тем в литературе даже не ставилась, вплоть до появления больших языковых моделей.

Суммаризация тем с использованием LLM требует аккуратного предварительного отбора и ранжирования тематически релевантных словосочетаний, фраз, предложений. Поиск сбалансированного сочетания экстрактивных техник на базе РТМ с генеративными LLM представляется наиболее здравым подходом к суммаризации тем. Данная задача на сегодняшний день также не имеет эффективного общепринятого решения.

6. Проблема автоматического добавления новых тем и удаления старых в текстовых потоках. Задача обнаружения и отслеживания тем (topic detection and tracking, TDT) хорошо известна в литературе, начиная с отчёта DARPA [3], 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 терминами, 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 (новостей, патентных баз, научных публикаций, социальных медиа, электронной почты).

7. Проблема согласования тематических векторов с предобученными нейросетевыми эмбедингами. Нейросетевые LLM предобучены на больших данных, по сути, на всём текстовом материале, накопленном человечеством за всю свою историю. Естественно предполагать, что число различных тем растёт с ростом всего доступного объёма текстов. Каждое научное открытие, мировое событие, сюжетная линия достойны отдельной темы.

Для LLM новая тема — это новый кластер в пространстве эмбедингов. Кластеры появляются, но размерность пространства остаётся фиксированной. Для РТМ ситуация принципиально иная. Каждая новая тема увеличивает на единицу размерность всех тематических векторов в модели. Это не катастрофа, если тематика коллекции

относительно стабильна или растёт медленно с ростом коллекции. Это становится катастрофой, если подумать о предобучении тематической модели на всём Интернете. Преимущество интерпретируемости и разреженности вероятностных тематических векторов оборачивается слабостью. Ниша тематического моделирования оказывается ограниченной текстовыми коллекциями с фиксированным или хотя бы медленно растущим числом тем.

Нейросетевые тематические модели (neural topic model, NTM), такие как BERTopic [42] или Contextual-Top2Vec [4], отчасти преодолевают это ограничение. Они используют предобученную LLM, чтобы построить по входной коллекции множество эмбедингов. Затем это множество разбивается алгоритмом кластеризации на фиксированное число тематически различающихся кластеров. Слова или фрагменты документов относятся к ближайшим кластерам. В результате каждый документ получает распределение по темам $p(t|d)$, а каждая тема образуется фразами, а не словами. Это упрощает автоматическое именование и суммаризацию тем. Описанная стратегия реализуется конвейером (pipeline) из примерно десятка алгоритмов, включая LLM, агрегирование эмбедингов, понижение их размерности и кластеризацию, оптимизацию числа тем, токенизацию и сегментацию текстов, формирование тематических распределений, именование и суммаризацию тем.

Аналогичный сценарий возможен в ARTM на основе модели локальных контекстов, причём большинство шагов конвейера либо встраиваются в ЕМ-алгоритм, либо становятся лишними. Открытой проблемой остаётся, главным образом, первый шаг — инициализация тематической модели с помощью предобученной LLM. Это особенно важно для коллекций малого объёма, когда данных не хватает для статистически надёжного оценивания тем. В таких случаях LLM способна компенсировать нехватку данных.

8. Проблема автоматического подбора гиперпараметров, особенно в режиме онлайн-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 ЕМ-алгоритма: число итераций, размер пакетов, темп забывания предыдущих пакетов, число пакетов для обновления матрицы Φ . Применение генетических алгоритмов или методов AutoML приводит к громоздким и вычислительно трудоёмким алгоритмам. Представляется перспективным многокритериальное оценивание качества модели по пакетам с использованием методов планирования экспериментов.

12 Заключение

Вероятностное тематическое моделирование развивается как область исследований уже четверть века. За это время разработаны сотни моделей, опубликованы тысячи исследовательских работ. Однако подавляющее большинство из них так и остались «моделью для одной статьи». На практике многие исследователи предпочитают не замечать этого разнообразия и используют, главным образом, морально устаревшую модель LDA, которая имеет слабые механизмы адаптации под задачу и может давать посредственные результаты.

Возможно, этот парадокс объясняется тем, что в рамках байесовского обучения так и не удалось выработать гибкую технологию для конструирования тематических моделей с требуемыми свойствами. Такая технология должна была бы включать в себя библиотеку взаимозаменяемых моделей с унифицированным программным интерфейсом, чтобы их было легко оценивать, выбирать и комбинировать.

Практика применения тематического моделирования показывает, что к тематической модели могут предъявляться разнообразные сочетания требований из следующего списка, не претендующего на полноту:

- (а) *интерпретируемость*: каждая тема понятна экспертам и «умеет рассказать о себе», т. е. имеет название и суммаризацию, генерируемые автоматически;
- (б) *терминологичность*: лексическое ядро каждой темы образуется не только словами, но и словосочетаниями (n -граммами, коллокациями, терминами, номенами), что существенно улучшает интерпретируемость тем;
- (в) *разреженность*: каждый документ содержит небольшое число тем; каждая тема имеет лексическое ядро, существенно отличающее её от других тем;
- (г) *иерархичность*: темы делятся на подтемы иерархически, отражая тематическую структуру предметной области текстовой коллекции;
- (д) *мультиязычность*: каждая тема имеет своё представление в каждом языке;
- (е) *мультимодальность*: документы могут содержать метаданные и нетекстовые токены (авторы, источники, время, категории, ссылки, тэги и т. д.);
- (ж) *транзакционность*: данные могут содержать записи о взаимодействии объектов различной природы, не только парных, но и n -арных;
- (з) *темпоральность*: темы могут появляться, исчезать и меняться со временем;
- (и) *сегментность*: текст может состоять из тематически однородных сегментов;
- (к) *обучаемость* по пользовательской разметке, включая ручную коррекцию тем;
- (л) *адаптивность*: алгоритм обучения настраивает все эвристики и гиперпараметры модели под входную коллекцию автоматически, без участия пользователя;
- (м) *масштабируемость*: ЕМ-алгоритм для обработки больших данных должен быть онлайн-овый, параллельный, поблочный, возможно, распределённый.

Практически все перечисленные выше требования были реализованы в своё время по отдельности, в различных тематических моделях, главным образом, в рамках байесовского подхода. К сожалению, совмещение произвольного подмножества требований в одной модели является для байесовского обучения трудным вызовом.

Теория аддитивной регуляризации тематических моделей (ARTM) была разработана для того, чтобы решить именно эту проблему [94]. Каждому требованию соответствует оптимизационный критерий, называемый регуляризатором. Совместить несколько требований в одной модели — значит добавить сумму регуляризаторов к основному оптимизационному критерию $\log$ -правдоподобия.

ARTM существенно упрощает математические выкладки по сравнению с байесовским обучением. Фактически, ARTM превращает вероятностное тематическое моделирование из полигона для отработки навыков байесовского вывода в изящную теорию, вытекающую из одной общей леммы.

Библиотека **BigARTM** стала первым открытым «лего-конструктором» для тематического моделирования [98]. Модель с желаемым сочетанием требований собирается из готовых модулей-регуляризаторов. Большинство из перечисленных выше требований были переведены из байесовского формализма в ARTM и реализованы в **BigARTM**, что отражено в данном обзоре. Наиболее полное и последовательное описание ARTM и **BigARTM** можно найти в книге [97].

Практическим затруднением пока остаётся подбор коэффициентов регуляризации и других гиперпараметров. Это требует понимания математической теории ARTM, навыков программирования и опыта в дизайне экспериментов. Для снижения барьеров входа были разработаны фреймворки **TopicNet** [19] и **AutoTM 2.0** [56]. Они расширяют функциональность **BigARTM** в отдельных аспектах, но часть полезных функций, наоборот, не выносят на уровень интерфейса, тем самым скрывая от пользователя. Реализованные в них сценарии использования тематического моделирования всё ещё далеки от потребностей конечных пользователей (в частности, специалистов в области социо-гуманитарных исследований), и соответствуют, скорее, интересам разработчиков, которым важно продемонстрировать возможности новых моделей и их технологические преимущества. Создание гибкой и доступной среды для конечного пользователя на основе теории ARTM пока остаётся открытой проблемой.

В эпоху больших языковых моделей PTM не теряет актуальности. Тематические модели занимают нишу кластерного анализа больших текстовых коллекций, обеспечивая скорость, полноту и интерпретируемость тем. Никакая LLM на сегодняшний день не может гарантировать полноту разложения коллекции по темам, особенно когда коллекция не помещается в контекстное окно.

При этом LLM способны дополнить PTM в двух аспектах.

Во-первых, инициализация тем с использованием предобученных нейросетевых эмбедингов может способствовать образованию более адекватных тем, особенно на малых данных.

Во-вторых, суммаризация тем с помощью LLM на основе экстракции не только слов, но и n -грамм, фраз и фрагментов тематической моделью, может способствовать лучшему пониманию каждой темы пользователями.

Нейросетевые тематические модели (NTM) образуют новый быстро растущий тренд в области тематического моделирования. Они плохо совместимы с байесовскими вероятностными тематическими моделями на уровне постановок задач, поэтому гибридизация этих двух подходов затруднена. Оптимизационная техника ARTM гораздо ближе к обучению нейронных сетей, что позволяет надеяться на плодотворное взаимное обогащение.

Обзор подготовлен при финансовой поддержке Министерства экономического развития Российской Федерации (соглашение № 139-15-2025-012 от 20.06.2025, ИГК 000000Ц313925Р4Н0002).

При подготовке текста данного обзора ни одна LLM не использовалась.

Список литературы

- [1] *Alekseev V., Egorov E., Vorontsov K., Goncharov A., Nurumov K., Buldybayev T.* TopicBank: Collection of coherent topics using multiple model training with their further use for topic model validation // *Data and Knowledge Engineering*. — 2021. — Vol. 135, no. 101921.
- [2] *Alekseev V. A., Bulatov V. G., Vorontsov K. V.* Intra-text coherence as a measure of topic models interpretability // *Computational Linguistics and Intellectual Technologies. Dialogue* 2018. — 2018. — Pp. 1–13.
- [3] *Allan J., Carbonell J., Doddington G., Yamron J., Yang Y.* Topic detection and tracking pilot study: Final report // *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*. — Lansdowne, VA, USA: 1998. — Pp. 194–218.
- [4] *Angelov D., Inkpen D.* Topic modeling: Contextual token embeddings are all you need // *Findings of the Association for Computational Linguistics: EMNLP 2024* / Ed. by Y. Al-Onaizan, M. Bansal, Y.-N. Chen. — Association for Computational Linguistics, 2024. — Pp. 13528–13539.
- [5] *Apishev M., Koltcov S., Koltsova O., Nikolenko S., Vorontsov K.* Additive regularization for topic modeling in sociological studies of user-generated text content // *15th Mexican International Conference on Artificial Intelligence*. — Vol. 10061. — Springer, LNAI, 2016. — Pp. 166–181.
- [6] *Apishev M., Koltcov S., Koltsova O., Nikolenko S., Vorontsov K.* Mining ethnic content online with additively regularized topic models // *Computacion y Sistemas*. — 2016. — Vol. 20, no. 3. — Pp. 387–403.
- [7] *Apishev M. A.* Effective implementations of topic modeling algorithms // *Trudy ISP RAN*. — 2020. — Vol. 32, no. 1. — Pp. 137–152.
- [8] *Apishev M. A., Vorontsov K. V.* Learning topic models with arbitrary loss // *Proceeding of the 26th Conference of FRUCT (Finnish-Russian University Cooperation in Telecommunications) Association*. — 2020. — Pp. 30–37.
- [9] *Asuncion A., Welling M., Smyth P., Teh Y. W.* On smoothing and inference for topic models // *Proceedings of the International Conference on Uncertainty in Artificial Intelligence*. — 2009. — Pp. 27–34.
- [10] *Balikas G., Amini M., Clausel M.* On a topic model for sentences // *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*. — SIGIR '16. — New York, NY, USA: ACM, 2016. — Pp. 921–924.
- [11] *Belyy A. V., Seleznova M. S., Sholokhov A. K., Vorontsov K. V.* Quality evaluation and improvement for hierarchical topic modeling // *Computational Linguistics and Intellectual Technologies. Dialogue* 2018. — 2018. — Pp. 110–123.
- [12] *Blei D. M.* Probabilistic topic models // *Communications of the ACM*. — 2012. — Vol. 55, no. 4. — Pp. 77–84.

- [13] *Blei D. M., Ng A. Y., Jordan M. I.* Latent dirichlet allocation // *Advances in Neural Information Processing Systems 14*, NIPS 2001, December 3–8, 2001, Vancouver, British Columbia, Canada] / Ed. by T. G. Dietterich, S. Becker, Z. Ghahramani. — MIT Press, 2001. — Pp. 601–608.
- [14] *Blei D. M., Ng A. Y., Jordan M. I.* Latent Dirichlet allocation // *Journal of Machine Learning Research*. — 2003. — Vol. 3. — Pp. 993–1022.
- [15] *Bodrunova S., Koltsov S., Koltsova O., Nikolenko S. I., Shimorina A.* Interval semi-supervised LDA: Classifying needles in a haystack // *MICAI (1)* / Ed. by F. C. Espinoza, A. F. Gelbukh, M. Gonzalez-Mendoza. — Vol. 8265 of *Lecture Notes in Computer Science*. — Springer, 2013. — Pp. 265–274.
- [16] *Boyd-Graber J., Hu Y., Mimno D.* Applications of topic models // *Foundations and Trends in Information Retrieval*. — 2017. — Vol. 11, no. 2-3. — Pp. 143–296.
- [17] *Brown T. B., Mann B., Ryder N., Subbiah M., Kaplan J., Dhariwal P., Neelakantan A. et al.* Language models are few-shot learners // *Advances in Neural Information Processing Systems* / Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin. — Vol. 33. — Curran Associates, Inc., 2020. — Pp. 1877–1901.
- [18] *Bulatov V., Alekseev V., Vorontsov K.* Determination of the number of topics intrinsically: Is it possible? // *Recent Trends in Analysis of Images, Social Networks and Texts*, 11-th International Conference AIST-2023, Yerevan, Armenia, September 28–30, 2023. — Springer Nature Switzerland, 2024. — Pp. 3–17.
- [19] *Bulatov V., Egorov E., Veselova E., Polyudova D., Alekseev V., Goncharov A., Vorontsov K.* TopicNet: Making additive regularisation for topic modelling accessible // *Proceedings of The 12th Conference on Language Resources and Evaluation (LREC 2020)*. — 2020. — Pp. 6745–6752.
- [20] *Chang J., Gerrish S., Wang C., Boyd-Graber J. L., Blei D. M.* Reading tea leaves: How humans interpret topic models // *Neural Information Processing Systems (NIPS)*. — 2009. — Pp. 288–296.
- [21] *Chemudugunta C., Smyth P., Steyvers M.* Modeling general and specific aspects of documents with a probabilistic topic model // *Advances in Neural Information Processing Systems*. — Vol. 19. — MIT Press, 2007. — Pp. 241–248.
- [22] *Chirkova N. A., Vorontsov K. V.* Additive regularization for hierarchical multimodal topic modeling // *Journal Machine Learning and Data Analysis*. — 2016. — Vol. 2, no. 2. — Pp. 187–200.
- [23] *Churchill R., Singh L.* The evolution of topic modeling // *ACM Comput. Surv.* — 2022. — Vol. 54, no. 10s. — 35 pp.
- [24] *Dai A. M., Olah C., Le Q. V.* Document embedding with paragraph vectors // *NIPS Deep Learning Workshop*. — 2015.
- [25] *Datyev I., Fedorov A.* Additive regularization for topic modeling of social media communities (in russian) // *Ontology of Designing*. — 2022. — Vol. 12, no. 2(44). — Pp. 186–199.

- [26] *Daud A., Li J., Zhou L., Muhammad F.* Knowledge discovery through directed probabilistic topic models: a survey // *Frontiers of Computer Science in China*. — 2010. — Vol. 4, no. 2. — Pp. 280–301.
- [27] *Dauphin Y. N., Fan A., Auli M., Grangier D.* Language modeling with gated convolutional networks // *Proceedings of the 34th International Conference on Machine Learning – Volume 70*. — ICML'17. — JMLR.org, 2017. — Pp. 933–941.
- [28] *Devlin J., Chang M.-W., Lee K., Toutanova K.* BERT: Pre-training of deep bidirectional transformers for language understanding // *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*. — Minneapolis, Minnesota: ACL, 2019. — Pp. 4171–4186.
- [29] *Dudarenko M. A.* Regularization of multilingual topic models (in russian) // *Vychisl. Metody Programm. (Numerical methods and programming)*. — 2015. — Vol. 16. — Pp. 26–38.
- [30] *Egorov E., Nikitin F., Goncharov A., Alekseev V., Vorontsov K.* Topic modelling for extracting behavioral patterns from transactions data // *Proceedings of International Conference on Artificial Intelligence: Applications and Innovations, Vrdnik Banja, Serbia, 30 September – 4 October 2019*. — 2019. — Pp. 40–44.
- [31] *El-Kishky A., Song Y., Wang C., Voss C. R., Han J.* Scalable topical phrase mining from text corpora // *Proc. VLDB Endowment*. — 2014. — Vol. 8, no. 3. — Pp. 305–316.
- [32] *Fedorov A. M., Datyev I. O., Shchur A. L.* Social media communities topic modeling // *Data Science and Intelligent Systems* / Ed. by R. Silhavy, P. Silhavy, Z. Prokopova. — Cham: Springer International Publishing, 2021. — Pp. 605–614.
- [33] *Fedorov A. M., Datyev I. O., Vishnyakov I. G.* Designing an information system for integrated topic analysis of social media big data (in russian) // *Ontology of Designing*. — 2024. — Vol. 14, no. 1. — Pp. 55–70.
- [34] *Feldman D. G., Sadekova T. R., Vorontsov K. V.* Combining facts, semantic roles and sentiment lexicon in a generative model for opinion mining // *Computational Linguistics and Intellectual Technologies. Dialogue 2020*. — 2020. — Pp. 268–283.
- [35] *Frei O., Apishev M.* Parallel non-blocking deterministic algorithm for online topic modeling // *AIST'2016, Analysis of Images, Social networks and Texts*. — Vol. 661. — Springer International Publishing Switzerland, Communications in Computer and Information Science (CCIS), 2016. — Pp. 132–144.
- [36] *Gerasimenko N., Chernyavskiy A., Nikiforova M., Ianina A., K. V.* Incremental topic modeling for scientific trend topics extraction // *Computational Linguistics and Intellectual Technologies. Dialogue 2023*. — 2023. — Pp. 88–103.
- [37] *Gerasimenko N. A., Chernyavsky A. S., Nikiforova M. A., Nikitin M. D., Vorontsov K. V.* Incremental learning of topic models for finding trend topics

- in scientific publications // *Doklady Mathematics*. — 2022. — Vol. 106, no. 1. — Pp. S97–S98.
- [38] *Gerasimenko N. A., Potapova P. S., Ianina A. O., Vorontsov K. V.* Application of probabilistic topic modeling in four exploratory search problems (in russian) // *Newsletter of the Russian Library Association*. — 2022. — no. 98. — Pp. 43–48.
 - [39] *Girolami M., Kabán A.* On an equivalence between PLSI and LDA // SIGIR'03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval. — 2003. — Pp. 433–434.
 - [40] *Gorbulev A., Alekseev V., Vorontsov K.* Iterative improvement of an additively regularized topic model // Analysis of Images, Social Networks and Texts: 12th International Conference, AIST 2024, Bishkek, Kyrgyzstan, October 17–19, 2024, Revised Selected Papers. — Berlin, Heidelberg: Springer-Verlag, 2025. — Pp. 87–106.
 - [41] *Grant C. E., George C. P., Kanjilal V., Nirkhiwale S., Wilson J. N., Wang D. Z.* A topic-based search, visualization, and exploration system // FLAIRS Conference. — AAAI Press, 2015. — Pp. 43–48.
 - [42] *Grootendorst M.* BERTopic: Neural topic modeling with a class-based TF-IDF procedure. — arXiv preprint arXiv:2203.05794. — 2022.
 - [43] *Hofmann T.* Probabilistic latent semantic indexing // Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. — New York, NY, USA: ACM, 1999. — Pp. 50–57.
 - [44] *Hofmann T.* Unsupervised learning by probabilistic latent semantic analysis // *Machine Learning*. — 2001. — Vol. 42, no. 1-2. — Pp. 177–196.
 - [45] *Hofmann T., Puzicha J.* Latent class models for collaborative filtering // Proceedings of the 16th International Joint Conference on Artificial Intelligence - Volume 2. — IJCAI'99. — San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1999. — Pp. 688–693.
 - [46] *Ianina A., Golitsyn L., Vorontsov K.* Multi-objective topic modeling for exploratory search in tech news // Communications in Computer and Information Science, vol 789. AINL-6: Artificial Intelligence and Natural Language Conference, St. Petersburg, Russia, September 20-23, 2017 / Ed. by A. Filchenkov, L. Pivovarova, J. Žižka. — Springer International Publishing, Cham, 2018. — Pp. 181–193.
 - [47] *Ianina A., Vorontsov K.* Multimodal topic modeling for exploratory search in collective blog // *Machine Learning and Data Analysis*. — 2016. — Vol. 2, no. 2. — Pp. 173–186.
 - [48] *Ianina A., Vorontsov K.* Regularized multimodal hierarchical topic model for document-by-document exploratory search // Proceeding Of The 25th Conference Of FRUCT (Finnish-Russian University Cooperation in Telecommunications) Association. The seminar on Intelligence, Social Media and Web (ISMW). Helsinki, Finland, November 5–8, 2019. / Ed. by S. Balandin, V. Niemi, T. Tutina. — 2019. — Pp. 131–138.

- [49] *Ianina A. O., Vorontsov K. V.* Hierarchical interpretable topical embeddings for exploratory search and real-time document tracking // *International Journal of Embedded and Real-Time Communication Systems (IJERTCS)*. — 2020. — Vol. 11, no. 4. — 19 pp.
- [50] *Irkhin I. A., Bulatov V. G., Vorontsov K. V.* Additive regularization of topic models with fast text vectorization // *Computer Research and Modeling*. — 2020. — Vol. 12, no. 6. — Pp. 1515–1528.
- [51] *Irkhin I. A., Vorontsov K. V.* Convergence of the algorithm of additive regularization of topic models // *Trudy Instituta Matematiki i Mekhaniki UrO RAN*. — 2020. — Vol. 26, no. 3. — Pp. 56–68.
- [52] *Jähnichen P., Oesterling P., Heyer G., Liebmann T., Scheuermann G., Kuras C.* Exploratory search through visual analysis of topic models // *Proceedings of the Digital Humanities*. — 2015. — Vol. 6, no. 8. — P. 13.
- [53] *Jelodar H., Wang Y., Yuan C., Feng X., Jiang X., Li Y., Zhao L.* Latent dirichlet allocation (LDA) and topic modeling: models, applications, a survey // *Multimedia Tools and Applications*. — 2019. — Vol. 78, no. 11. — Pp. 15169–15211.
- [54] *Khodorchenko M., Butakov N., Nasonov D.* Improved evolutionary approach for tuning topic models with additive regularization // Hybrid Artificial Intelligent Systems: 18th International Conference, HAIS 2023, Salamanca, Spain, September 5–7, 2023, Proceedings. — Berlin, Heidelberg: Springer-Verlag, 2023. — Pp. 409–420.
- [55] *Khodorchenko M., Butakov N., Sokhin T., Teryoshkin S.* Surrogate-based optimization of learning strategies for additively regularized topic models // *Logic Journal of the IGPL*. — 2022.
- [56] *Khodorchenko M., Butakov N., Zuev M., Nasonov D.* AutoTM 2.0: Automatic topic modeling framework for documents analysis. — 2024.
- [57] *Khodorchenko M., Teryoshkin S., Sokhin T., Butakov N.* Optimization of learning strategies for ARTM-based topic models // Hybrid Artificial Intelligent Systems / Ed. by E. A. de la Cal, J. R. Villar Flecha, H. Quintián, E. Corchado. — Springer International Publishing, 2020. — Pp. 284–296.
- [58] *Khrylchenko K. Y., Vorontsov K. V.* Optimizing modality weights in topic models of transactional data (in russian) // *Automation and Remote Control*. — 2022. — no. 83. — Pp. 1908–1922.
- [59] *Kochedykov D. A., Apishev M. A., Golitsyn L. V., Vorontsov K. V.* Fast and modular regularized topic modelling // Proceeding of the 21st Conference Of FRUCT (Finnish-Russian University Cooperation in Telecommunications) Association. The seminar on Intelligence, Social Media and Web (ISMW). Helsinki, Finland, November 6–10, 2017. — IEEE, 2017. — Pp. 182–193.
- [60] *Koltcov S., Koltsova O., Nikolenko S.* Latent Dirichlet allocation: Stability and applications to studies of user-generated content // Proceedings of the 2014 ACM

- Conference on Web Science.— WebSci'14.— New York, NY, USA: ACM, 2014.— Pp. 161–165.
- [61] *Krasnov F.* Topic classification through topic modeling with additive regularization for collection of scientific papers // Proceedings of the 14th Central and Eastern European Software Engineering Conference Russia.— CEE-SECR '18.— New York, NY, USA: Association for Computing Machinery, 2018.
 - [62] *Krasnov F.* Evaluation of optimal number of topics of topic model: An approach based on the quality of clusters // *International Journal of Open Information Technologies*.— 2019.— Vol. 7, no. 2.— Pp. 8–15.
 - [63] *Krasnov F., Ushmaev O.* Exploration of hidden research directions in oil and gas industry via full text analysis of onepetro digital library // *International Journal of Open Information Technologies*.— 2018.— Vol. 6, no. 5.— Pp. 7–14.
 - [64] *Krasnov F. V., Shvartsman M. E., Dimentov A. V., Sen A. I.* A thematic coherence study of a bilingual corpus of articles on oil and gas research // *Automatic Documentation and Mathematical Linguistics*.— 2019.— Vol. 53, no. 3.— Pp. 138–142.
 - [65] *Markovtsev V., Kant E.* Topic modeling of public repositories at scale using names in source code.— 2017.
 - [66] *McAuliffe J. D., Blei D. M.* Supervised topic models // Advances in Neural Information Processing Systems 20 / Ed. by J. C. Platt, D. Koller, Y. Singer, S. T. Roweis.— Curran Associates, Inc., 2008.— Pp. 121–128.
 - [67] *Mei Q., Shen X., Zhai C.* Automatic labeling of multinomial topic models // Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.— New York, NY, USA: Association for Computing Machinery, 2007.— Pp. 490–499.
 - [68] *Mikolov T., Chen K., Corrado G., Dean J.* Efficient estimation of word representations in vector space // *CoRR*.— 2013.— Vol. abs/1301.3781.
 - [69] *Milkova M. A.* Patent-based import substitution analysis with additively regularized topic models // *CEUR Workshop Proceedings*.— 2021.— no. 2830.— Pp. 16–27.
 - [70] *Mukhamediev R. I., Yelis M., Yakunin K., Popova Y., Kuchin Y., Symagulov A., Yunicheva N. et al.* Exploring the health care system's representation in the media through hierarchical topic modeling // *Cogent Engineering*.— 2024.— Vol. 11, no. 1.— P. 2324614.
 - [71] *Newman D., Karimi S., Cavedon L.* External evaluation of topic models // Australasian Document Computing Symposium.— 2009.— Pp. 11–18.
 - [72] *Newman D., Lau J. H., Grieser K., Baldwin T.* Automatic evaluation of topic coherence // Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics.— HLT'10.— Stroudsburg, PA, USA: ACL, 2010.— Pp. 100–108.

- [73] *Newman D., Noh Y., Talley E., Karimi S., Baldwin T.* Evaluating topic models for digital libraries // Proceedings of the 10th annual Joint Conference on Digital libraries. — JCDL '10. — New York, NY, USA: ACM, 2010. — Pp. 215–224.
- [74] *Nikolenko S. I.* ARTM vs. LDA: an SVD extension case study // Supplementary Proceedings of AIST'2016, Analysis of Images, Social networks and Texts. — Vol. 661. — Springer International Publishing Switzerland, Communications in Computer and Information Science (CCIS), 2016. — Pp. 276–282.
- [75] *Popov A., Bulatov V., Polyudova D., Veselova E.* Unsupervised dialogue intent detection via hierarchical topic model // Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019). — Varna, Bulgaria: INCOMA Ltd., 2019. — Pp. 932–938.
- [76] *Potapenko A., Popov A., Vorontsov K.* Interpretable probabilistic embeddings: bridging the gap between topic models and neural networks // Communications in Computer and Information Science, vol 789. AINL-6: Artificial Intelligence and Natural Language Conference, St. Petersburg, Russia, September 20-23, 2017. — Springer, Cham, 2017. — Pp. 167–180.
- [77] *Potapenko A. A., Vorontsov K. V.* Robust PLSA performs better than LDA // 35th European Conference on Information Retrieval, ECIR-2013, Moscow, Russia, 24-27 March 2013. — Lecture Notes in Computer Science (LNCS), Springer Verlag-Germany, 2013. — Pp. 784–787.
- [78] *Prokofyeva E., Zaytsev R., Maltseva S. V.* Application of modern data analysis methods to cluster the clinical pathways in urban medical facilities // *2019 IEEE 21st Conference on Business Informatics (CBI)*. — 2019. — Vol. 01. — Pp. 75–83.
- [79] *Reisenbichler M., Reutterer T.* Topic modeling in marketing: recent advances and research opportunities // *Journal of Business Economics*. — 2019. — Vol. 89, no. 3. — Pp. 327–356.
- [80] *Rönnqvist S.* Exploratory topic modeling with distributional semantics // Advances in Intelligent Data Analysis XIV: 14th International Symposium, IDA 2015, Saint Etienne. France, October 22 -24, 2015. Proceedings / Ed. by E. Fromont, T. De Bie, M. van Leeuwen. — Springer International Publishing, 2015. — Pp. 241–252.
- [81] *Scherer M., von Landesberger T., Schreck T.* Topic modeling for search and exploration in multivariate research data repositories // Research and Advanced Technology for Digital Libraries: International Conference on Theory and Practice of Digital Libraries, TPDL 2013, Valletta, Malta, September 22-26, 2013. Proceedings / Ed. by T. Aalberg, C. Papatheodorou, M. Dobрева, G. Tsakonas, C. J. Farrugia. — Springer Berlin Heidelberg, 2013. — Pp. 370–373.
- [82] *Skachkov N. A., Vorontsov K. V.* Improving topic models with segmental structure of texts // Computational Linguistics and Intellectual Technologies. Dialogue 2018. — 2018. — Pp. 652–661.

- [83] *Sokolov E., Bogolubsky L.* Topic models regularization and initialization for regression problems // *Proceedings of the 2015 Workshop on Topic Models: Post-Processing and Applications.* — New York, NY, USA: ACM, 2015. — Pp. 21–27.
- [84] *Solovyev V. D., Solnyshkina M. I., Tutubalina E. V.* Topic modeling for text structure assessment: The case of russian academic texts // *Journal of Language and Education.* — 2023. — Vol. 9, no. 3. — Pp. 144–159.
- [85] *Sukhareva A.* Completeness, stability and interpretability of probabilistic topic models (in russian) // *Intelligent Systems: Theory and Applications.* — 2020. — Vol. 24, no. 1. — Pp. 55–69.
- [86] *Sukhareva A., Vorontsov K.* Building a complete set of topics of probabilistic topic models (in russian) // *Intelligent Systems: Theory and Applications.* — 2019. — Vol. 23, no. 4. — Pp. 7–23.
- [87] *Sytnik A., Chernikova P., Vorontsov K., Bazlutckaia M.* Digital publics and the ukraine dilemma: Topic modelling of the cumulative twitter discussion // *Social Computing and Social Media: 16th International Conference, SCSM 2024, Held as Part of the 26th HCI International Conference, HCII 2024, Washington, DC, USA, June 29–July 4, 2024, Proceedings, Part III.* — Berlin, Heidelberg: Springer-Verlag, 2024. — Pp. 190–207.
- [88] *Tan Y., Ou Z.* Topic-weak-correlated latent Dirichlet allocation // *7th International Symposium Chinese Spoken Language Processing (ISCSLP).* — 2010. — Pp. 224–228.
- [89] *Teh Y. W., Jordan M. I., Beal M. J., Blei D. M.* Hierarchical Dirichlet processes // *Journal of the American Statistical Association.* — 2006. — Vol. 101, no. 476. — Pp. 1566–1581.
- [90] *Uteuov A., Kalyuzhnaya A.* Combined document embedding and hierarchical topic model for social media texts analysis // *Procedia Computer Science.* — 2018. — Vol. 136. — Pp. 293–303.
- [91] *Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser L., Polosukhin I.* Attention is all you need // *Advances in Neural Information Processing Systems 30* / Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett. — Curran Associates, Inc., 2017. — Pp. 5998–6008.
- [92] *Veselova E., Vorontsov K.* Topic balancing with additive regularization of topic models // *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, July 5–10, 2020.* — 2020. — Pp. 59–65.
- [93] *Vorontsov K., Frei O., Apishev M., Romov P., Suvorova M., Ianina A.* Non-bayesian additive regularization for multimodal topic modeling of large collections // *Proceedings of the 2015 Workshop on Topic Models: Post-Processing and Applications.* — New York, NY, USA: ACM, 2015. — Pp. 29–37.

- [94] Vorontsov K. V. Additive regularization for topic models of text collections // *Doklady Mathematics*. — 2014. — Vol. 89, no. 3. — Pp. 301–304.
- [95] Vorontsov K. V. A survey of probabilistic topic modeling (in russian) // *Automatic Natural Language Processing and Data Analysis: A Tutorial* / Ed. by E. I. Bolshakova, K. V. Vorontsov, M. I. Ionov, E. S. Klyshinsky, N. V. Lukashevich. — Moscow: HSE Publishing House, 2017. — 272 pp.
- [96] Vorontsov K. V. Rethinking probabilistic topic modeling from the point of view of classical non-bayesian regularization // *Data Analysis and Optimization*. — Springer, 2023. — Pp. 397–422.
- [97] Vorontsov K. V. Probabilistic Topic Modeling. ARTM Regularization Theory and the BigARTM Open Source Library (in Russian). Series: Textbook of the School of Applied Mathematics and Computer Science, MIPT. — URSS, 2025. — 224 pp.
- [98] Vorontsov K. V., Frei O. I., Apishev M. A., Romov P. A., Suvorova M. A. BigARTM: Open source library for regularized multimodal topic modeling of large collections // *AIST’2015, Analysis of Images, Social networks and Texts*. — Springer International Publishing Switzerland, Communications in Computer and Information Science (CCIS), 2015. — Pp. 370–384.
- [99] Vorontsov K. V., Frey A. I., Romov P. A., Yanina A. O., Suvorova M. A., Apishev M. A. BigARTM: An open source library for topic modeling of large text collections (in russian) // *Data Analytics and Management in Data-Intensive Domains. XVII International Conference DAMDID/RCDL’2015*. — NRNU MEPhI Obninsk, 2015. — Pp. 28–36.
- [100] Vorontsov K. V., Potapenko A. A. Regularization, robustness and sparsity of probabilistic topic models // *Computer Research and Modeling*. — 2012. — Vol. 4, no. 4. — Pp. 693–706.
- [101] Vorontsov K. V., Potapenko A. A. EM-like algorithms for probabilistic topic modeling // *Machine Learning and Data Analysis*. — 2013. — Vol. 1, no. 6. — Pp. 657–686.
- [102] Vorontsov K. V., Potapenko A. A. Regularization of probabilistic topic models to improve interpretability and determine the number of topics (in russian) // *International Conference on Computational Linguistics and Intellectual Technologies “Dialogue” (Bekasovo, June 4–8, 2014)*. — Issue 13 (20). — Moscow: RSUH Publishing House, 2014. — Pp. 676–687.
- [103] Vorontsov K. V., Potapenko A. A. Tutorial on probabilistic topic modeling: Additive regularization for stochastic matrix factorization // *AIST’2014, Analysis of Images, Social networks and Texts*. — Vol. 436. — Springer International Publishing Switzerland, Communications in Computer and Information Science (CCIS), 2014. — Pp. 29–46.
- [104] Vorontsov K. V., Potapenko A. A. Additive regularization of topic models // *Machine Learning, Special Issue on Data Analysis and Intelligent Optimization with Applications*. — 2015. — Vol. 101, no. 1. — Pp. 303–323.

- [105] *Vorontsov K. V., Potapenko A. A., Plavin A. V.* Additive regularization of topic models for topic selection and sparse factorization // The Third International Symposium On Learning And Data Sciences (SLDS 2015). April 20-22, 2015. Royal Holloway, University of London, UK. / Ed. by A. G. et al. — Springer International Publishing Switzerland 2015, 2015. — Pp. 193–202.
- [106] *Wallach H., Mimno D., McCallum A.* Rethinking LDA: Why priors matter // 23rd Annual Conference on Neural Information Processing Systems, Vancouver, BC, Canada / Ed. by Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, A. Culotta. — 2009. — Pp. 1973–1981.
- [107] *Wesslen R.* Computer-assisted text analysis for social science: Topic models and beyond // *CoRR*. — 2018. — Vol. abs/1803.11045.
- [108] *Yakunin K., Mukhamediev R. I., Yelis M., Symagulov A., Kuchin Y., Muhamedijeve E., Rabcan J., Margulan A.* Reflection of the COVID-19 pandemic in mass media // 2021 International Conference on Information and Digital Technologies (IDT). — Los Alamitos, CA, USA: IEEE Computer Society, 2021. — Pp. 260–263.
- [109] *Yakunin K., Mukhamediev R. I., Zaitseva E., Levashenko V., Yelis M., Symagulov A., Kuchin Y. et al.* Mass media as a mirror of the COVID-19 pandemic // *Computation*. — 2021. — Vol. 9, no. 12.
- [110] *Zavitsanos E., Paliouras G., Vouros G. A.* Non-parametric estimation of topic hierarchies from texts with hierarchical Dirichlet processes // *Journal of Machine Learning Research*. — 2011. — Vol. 12. — Pp. 2749–2775.
- [111] *Zhao H., Phung D., Huynh V., Jin Y., Du L., Buntine W.* Topic modelling meets deep neural networks: A survey // Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21 / Ed. by Z.-H. Zhou. — International Joint Conferences on Artificial Intelligence Organization, 2021. — Pp. 4713–4720.
- [112] *Zhao W. X., Jiang J., Weng J., He J., Lim E.-P., Yan H., Li X.* Comparing Twitter and traditional media using topic models // Proceedings of the 33rd European Conference on Advances in Information Retrieval. — ECIR'11. — Berlin, Heidelberg: Springer-Verlag, 2011. — Pp. 338–349.