На правах рукописи

Ирхин Илья Александрович

Единственность матричного разложения и сходимость регуляризованных алгоритмов в вероятностном тематическом моделировании

Специальность 05.13.17 —
«Теоретические основы информатики»

Автореферат
диссертации на соискание учёной степени
кандидата физико-математических наук

Москва — 2020

Работа выполнена в Московском физико-техническом институте (НИУ).

Научный руководитель: **Воронцов Константин Вячеславович**
доктор физико-математических наук, профессор РАН, ФГОАУ ВО «Московский физико-технический институт (НИУ)», руководитель лаборатории машинного интеллекта

Официальные оппоненты: **Елизаров Александр Михайлович,**
доктор физико-математических наук, профессор, ФГОАУ ВО «Казанский (Приволжский) федеральный университет», Высшая школа информационных технологий и интеллектуальных систем (ВШ ИТИС КФУ), профессор кафедры программной инженерии

Андрей Александрович Фильченков,
кандидат физико-математических наук, ФГОАУ ВО «Национальный исследовательский университет ИТМО», доцент факультета «Информационных технологий и программирования»

Ведущая организация: Федеральное государственное бюджетное учреждение науки Институт системного программирования им. В.П. Иванникова Российской академии наук

Защита состоится 24 декабря 2020 г. в 15 часов на заседании диссертационного совета Д 002.073.05 при Федеральном исследовательском центре «Информатика и управление» Российской Академии Наук по адресу: 119333, Москва, ул. Вавилова, д.40.

С диссертацией можно ознакомиться в библиотеке и на сайте ФИЦ ИУ РАН http://www.frccsc.ru.

Автореферат разослан ноября 2020 года.

Ученый секретарь
диссертационного совета
Д 002.073.05,
канд. техн. наук Рейер Иван Александрович

Общая характеристика работы

Актуальность темы. Развитие современных технологий сбора и хранения информации привело к заметному увеличению объёмов данных, в частности текстовых документов. Во многих прикладных областях возникает потребность в обработке и анализе накопленных текстовых коллекций. Одним из популярных в настоящее время направлений обработки естественного языка (Natural Language Processing, NLP) является тематическое моделирование. Тематическая модель описывает зависимость текстовых документов и содержащихся в них термов через наборы кластеров термов — тем. В роли термов обычно рассматриваются слова или их нормальные формы, но также иногда используются словосочетания или термины. Конкретная форма термов зависит то того, какие виды предварительной обработки текста были применены к коллекции. Тематическая модель для текстовой коллекции относит каждый документ к некоторым темам и для каждой темы определяет какие термы её образуют. Выявление подобных тематики текста можно рассматривать как шаг в направлении понимания естественного языка (Natural Language Understanding, NLU). В частности, тематическое моделирование предоставляет вариант решения проблемы синонимии и полисемии слов. Синонимы объединяются в одну тему, поскольку обычно употребляются в схожих контекстах. В то же время слова с несколькими значениями и омонимы попадают сразу в несколько тем, позволяя отличить разные по смыслу употребления друг от друга.

Тематическое моделирование может быть использовано для получения интерпретируемых векторных представлений слов, демонстрирующих сравнимое качество с векторными представлениями модели SGNS (Skip-Gram Negative Sampling) [*Distributed Representations of Words and Phrases and their Compositionality*, 2013] на задачах сравнения семантически близких слов [Potapenko, Popov, Vorontsov, 2017]. Но применение тематического моделирования не ограничивается только областью анализа текстов. Данный подход применяется и в других областях, например, в анализе аудио [W. Wang, 2011], анализе изображений и видео [Feng, Lapata, Mirella, 2010; Hospedales, Gong, Xiang, 2011; LI (и др.), 2012], биоинформатике [Pritchard J. K., 2000; Shivashankar (и др.), 2011]. Также тематические модели используются в задачах информационного поиска [Vulić, Smet, Moens, 2012; Vulić (и др.), 2015; Ianina, Golitsyn, Vorontsov, 2017; Ianina, Vorontsov, 2019] и рекомендаций [Nikolenko, 2015; Nikolenko, Koltcov, Koltsova, 2017; Pan, Li, 2010].

Общий подход для решения задачи тематического моделирования — построение вероятностной тематическая модели (Probabilistic Topic Model, PTM). Согласно этому подходу документы описываются некоторым дискретным распределением вероятностей на множестве тем, а темы —

дискретным распределением вероятностей на множестве термов. Построенная модель позволяет преобразовать любой текст в вектор вероятностей тем. Важным преимуществом тематического векторного представления текста является его интерпретируемость. Каждая координата вектора показывает долю соответствующей темы в тексте, при этом семантика темы описывается частотным словарём термов, то есть фактически словами естественного языка.

Классическим методом построения РТМ является предложенный в 1999 году вероятностный латентный семантический анализ (Probabilistic Latent Semantic Analysis, PLSA [Hofmann, 1999]). Этот подход задаёт вероятностную модель порождения термов в документах и строит разбиение термов и документов на темы, исходя из принципа максимизации правдоподобия. В 2003 году была предложена классическая модель латентного размещения Дирихле (Latent Dirichlet Allocation, LDA [Blei, Ng, Jordan, 2003]), которая была более устойчива и более точно учитывала данные о редких термах.

Важным преимуществом модели LDA является возможность расширять вероятностную модель дополнительными параметрам. Это сыграло важную роль в популярности подхода LDA [*Applications of topic models*, 2017]. Именно на основе этого подхода, были предложены вероятностные модели, учитывающие связи между документами [Cohn, Hofmann, 2001; McCallum, Corrada-Emmanuel, X. Wang, 2005; Nallapati, Cohen, 2008] или метаданные о документах [*Probabilistic author-topic models for information discovery*, 2004]. Также есть модели, которые учитывают время появления документа и его язык [Zosa, Granroth-Wilding, 2019] или порядок слов в документе [Gruber, Weiss, Rosen-Zvi, 2007; Wallach, 2006], что изначально было несвойственно подходу LDA.

Традиционный способ построения новых тематических моделей описан в [*Applications of topic models*, 2017]. Рекомендации включают в себя: введение новой вероятностной модели коллекции документов, которая не должна быть слишком вычислительно сложной, оставаясь при этом реалистичной; нахождение нового алгоритма оценки апостериорного распределения параметров; реализацию этого алгоритма; валидацию результатов. Трудностями использования подобного подхода являются необходимость проделывать данные действия заново для каждой новой модели, а также сложность и, иногда, невозможность построения тематических моделей, удовлетворяющих нескольким различным требований одновременно.

Теория аддитивной регуляризации тематических моделей (Additive Regularization of Topic Models, ARTM) [Vorontsov, Potapenko, 2015] решает эти проблемы, отказываясь от использования байесовского вывода [Vorontsov, Potapenko, 2014a]. В ARTM любые требования к модели формализуются через оптимизационные критерии — регуляризаторы.

Если требований несколько, то в постановку оптимизационной задачи вводится взвешенная сумма регуляризаторов [Vorontsov, Potapenko, 2015; Vorontsov, Potapenko, Plavin, 2015]. Байесовские тематические модели, как правило, удаётся переформулировать в терминах регуляризации, при этом существенно сокращается объём необходимых математических выкладок [*Fast and modular regularized topic modelling*, 2017]. Для оценивания параметров модели с произвольным набором регуляризаторов используется один и тот же итерационный процесс, называемый регуляризованным EM-алгоритмом. Этот алгоритм даёт возможность добавлять и заменять регуляризаторы не только на уровне постановки задачи, но и на уровне алгоритма и его программного кода. Это приводит к модульной технологии тематического моделирования, которая реализована в проектах с открытым кодом `BigARTM` [*BigARTM: Open Source Library for Regularized Multimodal Topic Modeling of Large Collections*, 2015; Frei, Apishev, 2017] и `TopicNet` [*TopicNet: Making Additive Regularisation for Topic Modelling Accessible*, 2020].

Дополнительным обоснованием использования регуляризаторов является некорректность по Адамару [Hadamard, 1902] поставленной оптимизационной задачи максимизации правдоподобия. Согласно теории регуляризации А. Н. Тихонова [Tikhonov, Arsenin, 1977], добавление регуляризатора доопределяет решение задачи и делает его устойчивым.

До сих пор в теории ARTM оставались открытыми вопросы о сходимости регуляризованного EM-алгоритма и о влиянии регуляризаторов на сходимость. В литературе хорошо изучены свойства Generalized Expectation Maximization алгоритма (GEM [Dempster, Laird, Rubin, 1977]), для которого известны достаточные условия сходимости [Wu, 1983]. В данной работе показывается, что итерации регуляризованного EM-алгоритма ARTM возможно интерпретировать как итерации GEM-алгоритма, за счёт чего возможно получить достаточные условия сходимости.

Для сходящегося итерационного процесса ARTM ставится вопрос о свойствах точки, к которой сошёлся алгоритм, например, открытым является вопрос о единственности полученного решения. С формальной точки зрения, в алгоритме ARTM для матрицы частот слов в документах строится стохастическое матричное разложение, ранг которого равен числу тем. Поскольку точного разложения требуемого ранга, как правило, не существует, то строится приближенное разложение, которое является локальным экстремумом оптимизируемого функционала. Таким образом, неединственность решения задачи тематического моделирования может возникать как из-за неоднозначности выбора этого приближения, так и из-за неединственности точного разложения приближения. До сих пор остаётся открытым вопрос о влиянии этих факторов на неединственность решения задачи тематического моделирования.

Проблема единственности стохастического матричного разложения исследовалась в работах [Donoho, Stodden, 2004; Laurberg (и др.), 2008; Gillis, 2012]. В этих работах представлены либо достаточные, либо необходимые условия единственности разложения. Недостатками предложенных условий применительно к тематическому моделированию являются их громоздкость и сложность проверки выполнения на практике.

В данной работе исследуются теоретические свойства регуляризованного EM-алгоритма ARTM. Особое внимание уделяются вопросам сходимости данного алгоритма и единственности стохастического матричного разложения в точке сходимости, поскольку они являются открытыми и представляют отдельный интерес. Также в рамках исследования производится поиск возможных модификаций алгоритма, которые за счёт теоретических гарантий будут улучшать качество получаемых тематических моделей.

Целью данной работы является получение достаточных условий сходимости алгоритма аддитивной регуляризации тематических моделей и достаточных условий для единственности стохастического матричного разложения в точке сходимости, которые могут быть проверены на реальных текстовых коллекциях, а также поиск модификаций исходного алгоритма, улучшающих сходимость и повышающих метрики качества тематических моделей.

Методология и методы исследования. В работе использованы подходы и методы численной оптимизации, вычислительной линейной алгебры, теории матричных разложений, машинного обучения. Для доказательства сходимости алгоритма ARTM использовались известные фундаментальные результаты о сходимости GEM-алгоритмов. Для доказательства единственности стохастического матричного разложения был использован подход геометрической интерпретации стохастического матричного разложения. В качестве реализации алгоритма ARTM использовались собственная реализация на языке Python[1] а также библиотеки с открытым кодом BigARTM [Frei, Apishev, 2017] и TopicNet [*TopicNet: Making Additive Regularisation for Topic Modelling Accessible*, 2020]. Для экспериментов в качестве текстовых коллекций использовались открытые публичные данные.

Научная новизна:

1. Впервые были получены достаточные условия сходимости алгоритма аддитивной регуляризации тематических моделей ARTM.
2. Были получены достаточные условия единственности стохастического матричного разложения в задачах тематического моделирования.
3. Были сформулированы причины нединственности решения для задач тематического моделирования.

[1] github.com/ilirhin/python_artm/tree/master/

4. Был разработан новый подход к стохастическому матричному разложению в тематическом моделировании, в котором одна из матриц находится в функциональной зависимости от другой.

Теоретическая значимость В работе впервые предложен подход с интерпретацией ARTM как GEM-алгоритма, в результате чего были получены достаточные условия сходимости данного алгоритма. Также были получены достаточные условия на единственность стохастического матричного разложения. В результате были сформулированы причины неединственности решения в тематическом моделировании.

Практическая значимость Разработана реализация алгоритма ARTM, с помощью которой теоретические положения диссертационной работы были подтверждены на реальных текстовых коллекциях. Предложенные в работе в работе алгоритмы реализованы в библиотеке с открытым кодом TopicNet. Модификации EM-алгоритма ARTM, полученные на основе теоретических результатов, значительно увеличивают основные метрики качества тематических моделей.

Основные положения, выносимые на защиту:

1. Теорема о достаточных условиях сходимости алгоритма ARTM.
2. Теорема о достаточных условиях единственности стохастического матричного разложения.
3. Модификация алгоритма ARTM, ускоряющая сходимость итерационного процесса.
4. Метод разреживания тематической модели, не увеличивающий перплексию получаемой модели.

Достоверность Достоверность результатов обеспечивается доказательствами теорем и описаниями проведённых экспериментов, допускающими их воспроизводимость, а также наличием репозитория Github с исходным кодом всех экспериментов.

Апробация работы. Основные результаты работы докладывались на:

1. 5th International Symposium, Conformal and Probabilistic Prediction with Applications, 2016
2. Научный семинар Школы Анализа Данных, 2016.
3. Научный семинар лаборатории искусственного интеллекта, 2018.
4. Научный семинар Федерального исследовательского центра «Информатика и Управление» Российской Академии Наук, 2020.

Личный вклад. Личный вклад диссертанта в работы, выполненные с соавторами, заключается в следующем:

– В работе [Селезнев, Ирхин, Кантор, 2018] предложена идея применения подхода тематического моделирования, предложены метрики качества, соответствующие решению прикладной задачи.

– В работе [Ирхин, Воронцов, 2020] предложены достаточные условия сходимости, доказаны все утверждения и теоремы, реализованы и проведены все эксперименты.
– В работе [Дербаносов, Ирхин, 2020] предложена и доказана основная лемма, реализованы и проведены все эксперименты.
– В работе [Ирхин, Булатов, Воронцов, 2020] предложена новая постановка оптимизационной задачи, выполнен вывод итераций алгоритма ARTM, реализована и проведена часть экспериментов, не связанная с библиотекой TopicNet.

Основные результаты по теме диссертации изложены в 3 печатных изданиях, 1 из которых изданы в журналах, рекомендованных ВАК, 2 — в периодических научных журналах, индексируемых Scopus.

Содержание работы

Во **введении** обосновывается актуальность исследований, проводимых в рамках данной диссертационной работы, приводится обзор научной литературы по изучаемой проблеме, формулируется цель, ставятся задачи работы, излагается научная новизна, теоретическая и практическая значимость представляемой работы.

В **первой главе** описывается постановка задачи тематического моделирования, вводятся основные определения и обозначения дисссертации. Далее приводятся оптимизационные задачи подходов PLSA и LDA и описывается алгоритм максимизациии выбранных функционалов. После чего описывается метод аддитивной регуляризации, ставится оптимизационная задача ARTM и приводится вывод стандартных формул оптимизирующего итерационного процесса.

Пусть D — конечное множество (коллекция) текстовых документов, W — конечное множество (словарь) всех употребляемых в них термов, T — конечное множество тем. Каждый документ $d \in D$ представляет собой последовательность n_d термов $(w_1, \ldots, w_{n_d})$ из словаря W. Принимается гипотеза «мешка слов», согласно которой порядок термов в документе не важен. Через n_{dw} обозначается число вхождений терма w в документ d.

Пусть $\phi_{wt} = p(w\,|\,t)$ — неизвестное распределение термов в темах, $\theta_{td} = p(t\,|\,d)$ — неизвестное распределения тем в документах. Задача вероятностного тематического моделирования заключается в том, чтобы найти параметры модели по эмпирическим данным n_{dw}. Для этого решается задача максимизации логарифма правдоподобия

$$L(\Phi,\Theta) = \sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{t\in T} \phi_{wt}\theta_{td} \;\to\; \max_{\Phi,\Theta} \qquad (1)$$

при ограничениях неотрицательности и нормировки:

$$\phi_{wt} \geq 0, \quad \sum_{w \in W} \phi_{wt} = 1, \qquad \theta_{td} \geq 0, \quad \sum_{t \in T} \theta_{td} = 1,$$

где Φ и Θ — матрицы параметров ϕ_{wt} и θ_{td} соответственно.

Задача 1 является некорректно поставленной по Адамару задачей приближённого стохастического матричного разложения $(\frac{n_{dw}}{n_d}) \approx \Phi\Theta$, имеющей в общем случае бесконечное множество решений. Чтобы выбрать из него наиболее подходящее решение, вводятся дополнительные критерии — регуляризаторы $R_i(\Phi,\Theta) \to \max$, $i = 1,\ldots,k$. В подходе ARTM предлагается максимизировать взвешенную сумму всех регуляризаторов $R(\Phi,\Theta) = \sum_{i=1}^{k} \tau_i R_i(\Phi,\Theta)$ совместно с основным критерием правдоподобия:

$$L(\Phi,\Theta) + R(\Phi,\Theta) = \sum_{d \in D} \sum_{w \in d} n_{dw} \log \sum_{t \in T} \phi_{wt}\theta_{td} + \sum_{i=1}^{k} \tau_i R_i(\Phi,\Theta) \;\to\; \max_{\Phi,\Theta}, \tag{2}$$

при тех же ограничениях неотрицательности и нормировки.

Наиболее известные тематические модели PLSA и LDA являются частными случаями регуляризации. В модели вероятностного латентного семантического анализа PLSA регуляризация не используется, $R(\Phi,\Theta) = 0$. В модели латентного размещения Дирихле LDA регуляризатором является логарифм правдоподобия априорного распределения Дирихле

$$R(\Phi,\Theta) = \sum_{t \in T} \sum_{w \in W} (\beta_w - 1) \ln \phi_{wt} + \sum_{d \in D} \sum_{t \in T} (\alpha_t - 1) \ln \theta_{td}$$

с гиперпараметрами β_w, α_t.

Применение теоремы Каруша–Куна–Таккера позволяет выписать систему уравнений для стационарных точек оптимизационной задачи (2). Решение данной системы методом простых итераций приводит к EM-подобному алгоритму, в котором на каждой итерации чередуются два шага: E-шаг (expectation) и M-шаг (maximization).

На E-шаге вычисляются значения условных вероятностей $p_{tdw} = p(t\,|\,d,w)$ по текущим значениям параметров ϕ_{wt} и θ_{td}:

$$p_{tdw} = \frac{\varphi_{wt}\theta_{td}}{\sum_s \varphi_{ws}\theta_{sd}}.$$

Данное выражение совпадает с формулой Байеса, поскольку, в силу гипотезы условной независимости, $p(t\,|\,d,w) = \frac{p(w\,|\,t)\;p(t\,|\,d)}{p(w\,|\,d)}$.

На M-шаге по условным вероятностям тем p_{tdw} для каждого терма в каждом документе вычисляются новые приближения параметров ϕ_{wt} и

θ_{td} и вспомогательные переменные n_{dwt}, n_{wt}, n_{td}, n_t, n_d, r_{wt}, r_{td}:

$$
\begin{aligned}
n_{dwt} &= n_{dw} p_{tdw}, \\
n_{wt} &= \sum_{d \in D} n_{dwt}, & n_{td} &= \sum_{w \in d} n_{dwt}, \\
n_t &= \sum_{w \in W} n_{wt}, & n_d &= \sum_{t \in T} n_{td}, \\
r_{wt} &= \phi_{wt} \frac{\partial R}{\partial \phi_{wt}}, & r_{td} &= \theta_{td} \frac{\partial R}{\partial \theta_{td}}, \\
\phi_{wt} &= \underset{w \in W}{\operatorname{norm}} \left(n_{wt} + r_{wt} \right), & \theta_{td} &= \underset{t \in T}{\operatorname{norm}} \left(n_{td} + r_{td} \right),
\end{aligned}
$$

где $\underset{i \in I}{\operatorname{norm}}(x_i) = \frac{(x_i)_+}{\sum_{j \in I}(x_j)_+}$ — операция нормировки, которая переводит произвольный числовой вектор $(x_i \colon i \in I)$ в дискретное вероятностное распределение, операция $(x_i)_+ = \max(x_i, 0)$ называется положительной срезкой.

Вспомогательные переменные n_* интерпретируются как оценки счётчиков: n_{dwt} — число вхождений терма w в документ d, связанных с темой t; n_{td} — число всех термов в документе d, связанных с темой t; n_{wt} — число раз, когда терм w был связан с темой t, во всей коллекции; n_t — число термов, связанных с темой t, во всей коллекции; n_d совпадает с длиной документа d.

Во **второй главе** рассматривается вопрос сходимости алгоритма ARTM. Итерации алгоритма ARTM рассматриваются как итерации GEM-алгоритма. Для этого вводится дополнительный функционал:

$$
Q(\Phi, \Theta, \Phi', \Theta') = \sum_{d,w,t} n_{dw} p'_{tdw} \ln(\phi_{wt} \theta_{td}) + R(\Phi, \Theta), \qquad \text{где } p'_{tdw} = \frac{\phi'_{wt} \theta'_{td}}{\sum\limits_t \phi'_{wt} \theta'_{td}}.
$$

Это стандартный приём при доказательстве сходимости GEM-алгоритма. Показывается, что изменения Q на итерациях являются нижней оценкой для изменений $L + R$:

$$
\Delta^k(L + R) \geq Q(\Phi^{k+1}, \Theta^{k+1}, \Phi^k, \Theta^k) - Q(\Phi^k, \Theta^k, \Phi^k, \Theta^k).
$$

Этот факт является основополагающим в доказательстве сходимости GEM-алгоритмов, предложенном в работе [Wu, 1983].

Далее вводятся требуемые определения, описывающие свойства регуляризаторов, которые нужны для сходимости алгоритма ARTM.

Определение 1. *Регуляризатор R является* δ-регулярным, *если на итерациях EM-алгоритма* $\forall t \ \exists w \colon n_{wt} + r_{wt} > \delta$ *и* $\forall d \ \exists t \colon n_{td} + r_{td} > \delta$. *Если регуляризатор обладает свойством δ-регулярности при некотором $\delta > 0$, то будем говорить, что регуляризатор сильно регулярен; при $\delta = 0$ будем просто говорить, что он регулярен.*

Регулярность гарантирует, что в операции norm не возникнет деления на нуль, то есть итерации корректно определены. Сильная же регулярность позволяет утверждать, что преобразования, которые производятся на итерациях алгоритма, являются непрерывными по (Φ,Θ). Это свойство легко выполняется на практике: если значение $n_{wt} + r_{wt}$ (или $n_{td} + r_{td}$) становится меньше δ, то вся тема (весь документ) исключается из модели и итерации продолжаются.

Определение 2. *Регуляризатор R* сохраняет нуль*, если на итерациях алгоритма из $n_{wt} = 0$ следует $\phi_{wt} = 0$ и из $n_{td} = 0$ следует $\theta_{td} = 0$.*

Это определение формализует следующие свойство итерационного процесса: если на какой-либо итерации значение ϕ_{wt} стало равным нулю, то оно будет оставаться нулевым на последующих итерациях, и аналогично для θ_{td}. Для регуляризатора данное свойство легко проверяется аналитически. На практике многие регуляризаторы им обладают. Регуляризатор модели LDA, на первый взгляд, не обладает данным свойством при $\beta_w > 1$ или $\alpha_t > 1$, так как при $n_{wt} = 0$ вполне может оказаться, что $\phi_{wt} > 0$. Однако при использовании ненулевой инициализации ϕ_{wt} значение n_{wt} не может обратиться в нуль. Поэтому и для такого регуляризатора условие сохранения нуля выполняется.

Определение 3. *Регуляризатор R называется* ϵ-разреживающим*, если на итерациях ЕМ-алгоритма $\phi_{wt}, \theta_{td} \notin (0, \epsilon)$.*

Некоторые регуляризаторы имеют неограниченную в окрестности нуля производную, поэтому при реализации ЕМ-алгоритма параметры, меньшие некоторого ϵ, зануляются. Это приводит к тому, что значения в матрице параметров оказываются отделены от нуля. Именно эта особенность отражена в данном определении.

Определение 4. *Регуляризатор R* корректный*, если на итерациях ЕМ-алгоритма из $n_{dw} > 0$ следует $p_{tdw} > 0$ хотя бы для одной темы t.*

Если модель даёт нулевую оценку вероятности $p(w\,|\,d) = 0$ при том, что терм w встречается в документе, $n_{dw} > 0$, то логарифм правдоподобия становится неограниченным, $L \to -\infty$. На практике этого легко избежать, если использовать регуляризатор сглаживания фоновых тем [Vorontsov, Potapenko, 2014b]. Он гарантирует, что для любого терма в любом документе найдётся хотя бы одна тема с ненулевой вероятностью.

В данных определениях формулируется основная теорема и следствия из неё:

Теорема 1. *Пусть регуляризатор R является дифференцируемой функцией при $\phi_{wt}, \theta_{td} \in (0, 1]$, сохраняющей нуль, корректной, ϵ-разрежижающей и δ-регулярной. Также допустим, что $Q(\Phi^{k+1}, \Theta^{k+1}, \Phi^k, \Theta^k) \geq$*

$Q(\Phi^k, \Theta^k, \Phi^k, \Theta^k)$ *начиная с некоторой итерации* k. *Тогда последовательность* p^k_{tdw} *сходится в смысле дивергенции Кульбака–Лейблера для любых* d *и* w *таких, что* $n_{dw} > 0$:

$$\mathrm{KL}\big(p^k_{tdw} \,\big\|\, p^{k+1}_{tdw}\big) \to 0 \text{ при } k \to \infty.$$

Следствие 1. *Если в дополнение к условиям Теоремы 1 регуляризатор* R *сильно регулярен, а* r_{wt} *и* r_{td} *непрерывны по* p_{tdw}, *то*

$$|\phi^k_{wt} - \phi^{k+1}_{wt}| \to 0 \text{ и } |\theta^k_{td} - \theta^{k+1}_{td}| \to 0.$$

Следствие 2. *Рассмотрим функцию* $F(\Phi, \Theta) = L(\Phi, \Theta) + R(\Phi, \Theta)$, *определённую для тех* Φ *и* Θ, *у которых множество нулевых позиций матриц совпадает с множеством ненулевых позиций* Ω, *стабилизировавшимся в ходе итераций.*

В условиях Следствия 1 если процесс не сошёлся в неподвижную точку, то все предельные точки траектории (Φ^k, Θ^k) *являются стационарными точками* F. *Если же множество стационарных точек* F *дискретно для каждого уровня значений* F, *то* (Φ^k, Θ^k) *сходится к некоторой стационарной точке* F.

Важным условием сходимости алгоритма ARTM является неуменьшение значения Q на M-шаге. Поэтому далее в работе производятся оценки изменения функционала Q на итерации. Доказывается теорема:

Теорема 2. *Пусть величины* r_{wt} *и* r_{td} *на M-шаге рассчитываются в точках*

$$\frac{n_{wt}}{\sum\limits_w n_{wt}} \text{ и } \frac{n_{td}}{\sum\limits_t n_{td}},$$

тогда в ходе нормировки $n_{wt} + r_{wt}$ *и* $n_{td} + r_{td}$ *на M-шаге при отсутствии занулений элементов матриц угол между вектором изменений* Δn *и градиентом* $\bar{R}$ *острый, если градиент ненулевой.*

На основе этой теоремы предлагается модификация исходного M-шага алгоритма, которая меняет расчёт величин r_{wt} и r_{td}. Формулы регуляризационных поправок из исходного алгоритма

$$r^k_{wt} = \phi^{k-1}_{wt} \frac{\partial R}{\partial \phi_{wt}}\big(\Phi^{k-1}_{wt}, \Theta^{k-1}_{td}\big); \qquad r^k_{td} = \theta^{k-1}_{td} \frac{\partial R}{\partial \theta_{td}}\big(\Phi^{k-1}_{wt}, \Theta^{k-1}_{td}\big); \tag{3}$$

заменяются на модифицированные согласно Теореме 2:

$$r^k_{wt} = \frac{n^k_{wt}}{n^k_t} \frac{\partial R}{\partial \phi_{wt}}\left(\frac{n^k_{wt}}{n^k_t}, \frac{n^k_{td}}{n^k_d}\right); \qquad r^k_{td} = \frac{n^k_{td}}{n^k_d} \frac{\partial R}{\partial \theta_{wt}}\left(\frac{n^k_{wt}}{n^k_t}, \frac{n^k_{td}}{n^k_d}\right). \tag{4}$$

В следующих разделах описываются и приводятся результаты экспериментов на текстовой коллекции, сравнивающих эти две формулы между собой на примере регуляризатора декорреляции:

$$R(\Phi) = -\frac{\tau}{|T|(|T|-1)} \sum_{t \neq s} \sum_{w \in W} \phi_{wt}\phi_{ws}.$$

Эксперименты показывают, что чем больше значение регуляризатора R по сравнению со значением логарифма правдоподобия L (с ростом коэффициента τ), тем больше положительный эффект от предложенной модификации (Рис. 1 и Таблица 1).

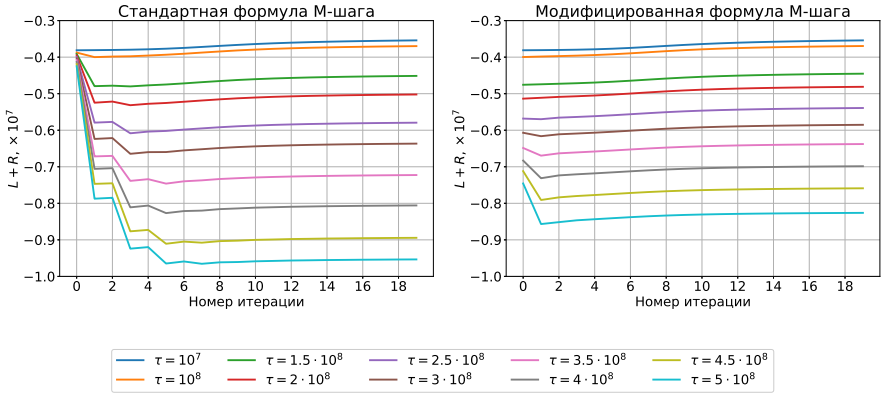

Рис. 1 — Изменение функционала $L + R$ на итерациях, $|T| = 30$, при различных значениях коэффициента регуляризации τ.

τ	$L + R$ стандарт	$L + R$ модификация	Улучшение $L + R$, %
10^7	-3536050	-3536340	-0.01
10^8	-3693905	-3691338	0.07
$1.5 \cdot 10^8$	-4509247	-4448501	1.35
$2.0 \cdot 10^8$	-5018335	-4808217	4.19
$2.5 \cdot 10^8$	-5790283	-5388187	6.94
$3.0 \cdot 10^8$	-6363392	-5848354	8.09
$3.5 \cdot 10^8$	-7223361	-6374974	11.75
$4.0 \cdot 10^8$	-8055262	-6982549	13.32
$4.5 \cdot 10^8$	-8941616	-7586618	15.15
$5.0 \cdot 10^8$	-9532948	-8259205	13.36

Таблица 1 — Итоговые значения $L + R$ по окончании итераций.

Также в экспериментах было подтверждено выполнение достаточных условий сходимости из Теоремы 1. Таким образом, было показано, что итерационный процесс ARTM сходится, после чего возникает вопрос анализа свойств точки, к которой сошёлся алгоритм. В частности, единственности полученного решения. Известно, что оптимизационная задача ARTM имеет неединственное решение, однако, причины этой неединственности подробно не изучались. Эффект неединственности потенциально состоит из двух частей: неединственности точного матричного разложения в точке, в которую сошёлся алгоритм, и мультиэкстремальности оптимизационной задачи.

В **третьей главе** определяется степень влияния этих двух факторов. Для этого рассматривается вопрос единственности точного стохастического матричного разложения. Сначала производится формальная постановка задачи, вводятся дополнительные обозначения.

Определение 5. *Матрица $F \in \mathbb{R}^{n\times m}$ будет называться неотрицательной, если все её элементы неотрицательны.*

Определение 6. *Неотрицательная матрица $F \in \mathbb{R}^{n\times m}$ будет называться стохастической, если $\forall j \sum_i F_{ij} = 1$.*

Определение 7. *Пусть дана матрица $F \in \mathbb{R}^{n\times m}$, её неотрицательным (стохастическим) матричным разложением будет называться представление в виде произведения $F = \Phi\Theta$ двух неотрицательных (стохастических) матриц $\Phi \in \mathbb{R}^{n\times k}$, $\Theta \in \mathbb{R}^{k\times m}$.*

Определение 8. *Пусть дана матрица $F \in \mathbb{R}^{n\times m}$, её матричным разложением полного ранга будет называться представление в виде произведения $F = \Phi\Theta$ двух матриц полного ранга $\Phi \in \mathbb{R}^{n\times k}$, $\Theta \in \mathbb{R}^{k\times m}$.*

Определение 9. *Разложение $F = \Phi\Theta$ будет называться единственным, если для любого другого разложения $F = \Phi'\Theta'$ выполняется $\Phi' = \Phi S, \Theta' = S^{-1}\Theta$, где S — некоторая матрица перестановки.*

$\overline{\mathrm{supp}(v)}$ — множество позиций, где стоят нулевые элементы вектора v;
$\mathrm{supp}(v)$ — множество позиций, где стоят ненулевые элементы вектора v;
X_j — j-ый столбец матрицы X;
$X\big[[i_1,\ldots,i_p],[j_1,\ldots,j_q]\big]$ — подматрица, состоящая из строк $i_1,\ldots,i_p$ и столбцов $j_1,\ldots,j_q$;

В этих обозначениях формулируется и доказывается теорема о достаточных условиях единственности матричного разложения:

Теорема 3. *Пусть дано разложение $F = \Phi\Theta$, $F \in \mathbb{R}^{n\times m}$, $\operatorname{rank} F = k$, $\Phi \in \mathbb{R}^{n\times k}$, $\Theta \in \mathbb{R}^{k\times m}$. Пусть выполнены условия:*

– $\forall i \in \{1,\ldots,k\}\ \exists j : \Theta_{ij} = 1, \forall i' \neq j\ \Theta_{i'j} = 0$;

– $\forall j\ \ \mathrm{rank}\Big(\Phi\big[\overline{\mathrm{supp}(\Phi_j)}, [1,\ldots,k] \setminus [j]\big]\Big) = k-1$.

Тогда разложение $F = \Phi\Theta$ единственно.

Далее приводится интерпретация с точки зрения тематического моделирования условия Теоремы 3.

Условие 1 требует наличия в матрице Θ единичной подматрицы размера $k \times k$. Матрица Θ отвечает за распределение тем в документах. Поэтому фактически это условие требует наличия в тематической модели k унитематических документов, то есть таких, в которых есть одна тема с вероятностью появления 1, а вероятности остальных тем нулевые. Выполнение этого условия можно гарантировать, добавив в коллекцию k искусственно созданных унитематических документов, слова для которых подбираются, например, экспертами.

Условие 2 говорит о том, что для любого j произведение матриц

$$\Phi\big[\overline{\mathrm{supp}(\Phi_j)}, [1,\ldots,k] \setminus [j]\big] \text{ и } \Theta\big[[1,\ldots,k] \setminus [j], :\big]$$

является неотрицательным матричным разложением полного ранга для матрицы

$$F\big[\overline{\mathrm{supp}(\Phi_j)}, :\big].$$

С точки зрения тематического моделирования это означает, что если для любой темы t из матрицы слова-документы F ранга T убрать все слова, встречающиеся в t-ой теме, то на получившей матрице слова-документы можно построить невырожденную тематическую модель на $T-1$ теме.

Далее описываются эксперименты на текстовой коллекции 20NewsGroups для проверки предложенных условий на практике. Предлагается эффективный способ проверки достаточных условий для матриц, полученных в итоге разложения. Чтобы для каждой темы t проверять полноту ранга матрицы $\Phi[\overline{supp(\Phi_t)}, [1,\ldots,T] \setminus [t]]$, находится минимальное сингулярное значение матрицы $\Phi[\overline{supp(\Phi_t)}, [1,\ldots,T] \setminus [t]]$ и сравнивается нулём. Это минимальное сингулярное значение для темы t обозначается σ_t . Далее описывается эксперимент и приводятся его результаты (Рис. 2), показывающие положительность σ_t на реальной текстовой коллекции.

Таким образом, показывается, что локальная неединственность оптимизационной задачи ARTM возникает из-за того, что разным значениям произведения $\Phi\Theta$ соответствуют одинаковые значения оптимизируемого функционала $L + R$. При этом для каждого значения $\Phi\Theta$ точное разложение на Φ и Θ определяется однозначно.

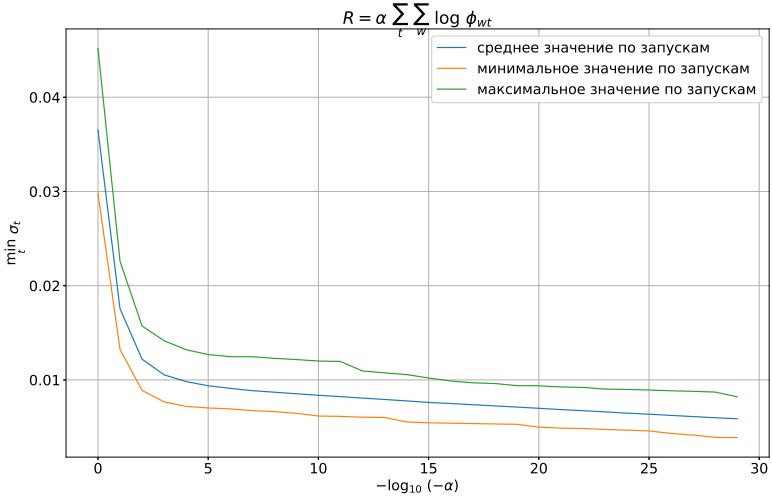

Рис. 2 — Изменение $\min\limits_t \sigma_t$ при стремлении коэффициента α к нулю в регуляризаторе $R(\Phi) = \alpha \sum\limits_t \sum\limits_w \ln \phi_{wt}$, т. е. при уменьшении силы разрежи-вания.

Важным фактором как сходимости, так и единственности точного матричного разложения является разреженность матрицы Ф. В **четвёртой главе** главе предлагается метод разреживания тематической модели без повышения её перплексии.

Доказывается теорема о изменении правдоподобия при зануления элементов матриц Ф и Θ:

Теорема 4. *Изменение значения L при зануления значения ϕ_{wt} составляет*

$$\Delta L = \sum_d n_{dw} \log(1 - p_{tdw}) + \sum_{d,u \neq w} n_{du} \log\left(1 + \frac{\phi_{wt}}{1 - \phi_{wt}} p_{tdu}\right).$$

Изменение значения L при зануления значения θ_{td} составляет

$$\Delta L = \sum_w n_{dw} \log\left(1 - p_{tdw}\right) - n_d \log(1 - \theta_{td}).$$

А также теорема об аппроксимации предложенных выражений:

Теорема 5. *Изменение значения L при занулении значения ϕ_{wt} после аппроксимации составляет*

$$\Delta L = \frac{n_t\phi_{wt} - n_{wt}}{1-\phi_{wt}} - \frac{1}{2}\left(\frac{\phi_{wt}}{1-\phi_{wt}}\right)^2 \sum_{d,u\neq w} n_{du}p_{tdu}^2 + \\ + \frac{1}{2}\sum_d n_{dw}p_{tdw}^2 + O\left(\sum_d n_{dw}p_{tdw}^3\right).$$

Изменение значения L при занулении значения θ_{td} после аппроксимации составляет

$$\Delta L = (n_d\theta_{td} - n_{td}) + \frac{1}{2}\sum_w n_{dw}p_{tdw}^2 - \frac{1}{2}n_d\theta_{td}^2 + O\left(\sum_w n_{dw}p_{tdw}^3\right).$$

На основе предложенных теорем предлагается метод разреживания тематических моделей. За γ_{wt} обозначается изменение L при занулении значения ϕ_{wt}. На итерациях EM-алгоритма ARTM зануляются элементы матрицы Φ, для которых $\gamma_{wt} \leq -\alpha$. Предложенная стратегия разреживания (OBD ARTM) сравнивается в эксперименте на коллекции 20Newsgroups с классическим методом разрежения с помощью регуляризатора разреживания (sparse ARTM), который соответствует занулению при $n_{wt} \leq \alpha$.

Метрика	Алгоритм	До	После	Увеличение, %
Разреженность	sparse ARTM	0.32	0.866	+**170**
Разреженность	OBD ARTM	0.32	0.86	+**168**
Перплексия	sparse ARTM	1518.1	2121.5	+**39**
Перплексия	OBD ARTM	1518.1	1549.8	+**2**

Таблица 2 — Разреженность и перплексия после 1 итерации разреживания разными методами

Результаты эксперимента (Таблица 2) показывает, что предложенный метод позволяет добиться примерно того же уровня разреженности, но при этом, в отличие от подхода с регуляризатором, не увеличивает перплексию модели.

Помимо разреженности во многих прикладных задачах важна интерпретируемость матрицы Φ, которая, обычно, оценивается через когерентность. Подход для анализа увеличения функционала Q, предложенный в первой главе, показывает как можно выводить формулы для M-шага, сохраняя при этом достаточные условия на сходимость. Используя этот подход, в **пятой главе** рассматривается изменение оптимизационной задачи ARTM, которое направлено на получение более разреженных и когерентных решений.

Предлагается заменить исходную оптимизационную задачу 2 на следующую:

$$L(\Phi, f(\Phi)) + R(\Phi, f(\Phi)) \to \max_{\Phi}, \tag{5}$$

где f — это некоторая функция, которая отображает матрицу темы-слова в матрицу документы-темы.

Подобное изменение мотивируется тем, что реализации тематического моделирования (особенно восстанавливающие элементы Θ "на лету") часто используют следующую эвристику: для получения θ_{td} конкретного документа d повторяются несколько итераций EM-алгоритма с фиксированной Φ. В этой процедуре вектор θ_{*d} сначала инициализируется некоторым образом (как правило, используется равномерное распределение), а затем итеративно обновляется по формуле $\theta_{td} \propto \sum_w n_{dw} p_{tdw}$ с пересчётом p_{tdw}. Обновление может происходить какое-то установленное количество итераций либо продолжаться до сходимости. То есть в реализациях тематического моделирования матрицы Φ и Θ находятся в некоторой функциональной зависимости, что и предлагается учесть в оптимизационной задаче.

Далее определяется функция зависимости Φ и Θ. В качестве интерпретируемой, простой для анализа и лёгкой для вычислений функции предлагается усреднение распределений тем слов по всем словам, встречающимся в документе. Более формально:

$$p(t \mid d) \propto \sum_w n_{dw} p(t \mid w),$$

где $p(t \mid w)$ получены по формуле Байеса, предполагая, что распределение $p(t)$ равномерно:

$$p(t \mid w) = \frac{p(w \mid t)}{\sum_{s=1}^{T} p(w \mid s)} = \frac{\Phi_{wt}}{\sum_s \Phi_{ws}}.$$

Откуда выводится:

$$\Theta_{td} = \sum_w \frac{n_{dw}}{\sum_u n_{du}} \frac{\Phi_{wt}}{\sum_s \Phi_{ws}}. \tag{6}$$

Далее приводится вывод формул E-шага и M-шага для 5 с функцией зависимости 6. Для этого вводятся дополнительные обозначения и доказывается теорема:

$$A_{dw} = \frac{n_{dw}}{\sum_s \Phi_{ws}\Theta_{sd}}[n_{dw} > 0], \quad B_{dw} = \frac{n_{dw}}{\sum_w n_{dw}},$$
$$C_{dt} = (A\Phi)_{dt} + \frac{\partial R}{\partial \Theta_{td}}, \quad h_w = \frac{1}{\sum_s \Phi_{ws}}.$$

Теорема 6. *В EM-алгоритме для 5 с формулой зависимости 6 E-шаг останется без изменений, а M-шаг будет выглядеть следующим образом:*

$$\Phi_{wt}^{new} \propto \left(\sum_d n_{dw} p_{tdw} + \Phi_{wt}^{old} \left(\frac{\partial R}{\partial \Phi_{wt}} + h_w (C^T B)_{tw} - h_w^2 (\Phi^{old} C^T B)_{ww} \right) \right)_+ \tag{7}$$

Для полученных формул проводится анализ, показывающий, что асимптотика времени работы новых формул не отличается от исходных. Также отмечается, что предложенные формул могут быть реализованы как регуляризатор для исходной постановки оптимизационной задачи:

$$r_{wt} = \Phi_{wt}^{old} \left(\frac{\partial R}{\partial \Phi_{wt}} + h_w (C^T B)_{tw} - h_w^2 (\Phi^{old} C^T B)_{ww} \right),$$

В последующих разделах описываются эксперименты и приводятся их результаты, показывающие улучшение основных метрик качества матрицы Φ по сравнению с PLSA и LDA (Рис. 3). Так демонстрируются результаты реализации предложенной модификации как регуляризатора в библиотеке для тематического моделирования TopicNet, которые показывают улучшение от использования модификации в комбинации с другими регуляризаторами (Таблица 3).

Алгоритм	Разреженность	Различность	PPMI	LogLift
sparse LDA	0.896	0.044	1.570	0.503
smooth LDA	0	0.043	1.509	0.479
PLSA	0.869	0.050	1.517	0.459
ARTM + *Reg*	0.898	0.027	1.710	0.590
TARTM	0.893	0.007	1.716	0.952
TARTM + *Reg*	**0.929**	**0.003**	**1.788**	**1.020**

Таблица 3 — Эксперимент с реализацией TopicNet. Сравнение моделей по четырём критериям: разреженность матрицы, различность тем, PPMI топ слов, LogLift. Модель TARTM достигает наилучших результатов по всем критериям кроме разреженности. Применение комбинации регуляризаторов сглаживания фоновых тем, разреживания предметных тем и декоррелирования (TARTM + *Reg*) существенно улучшает модель по всем пяти критериям.

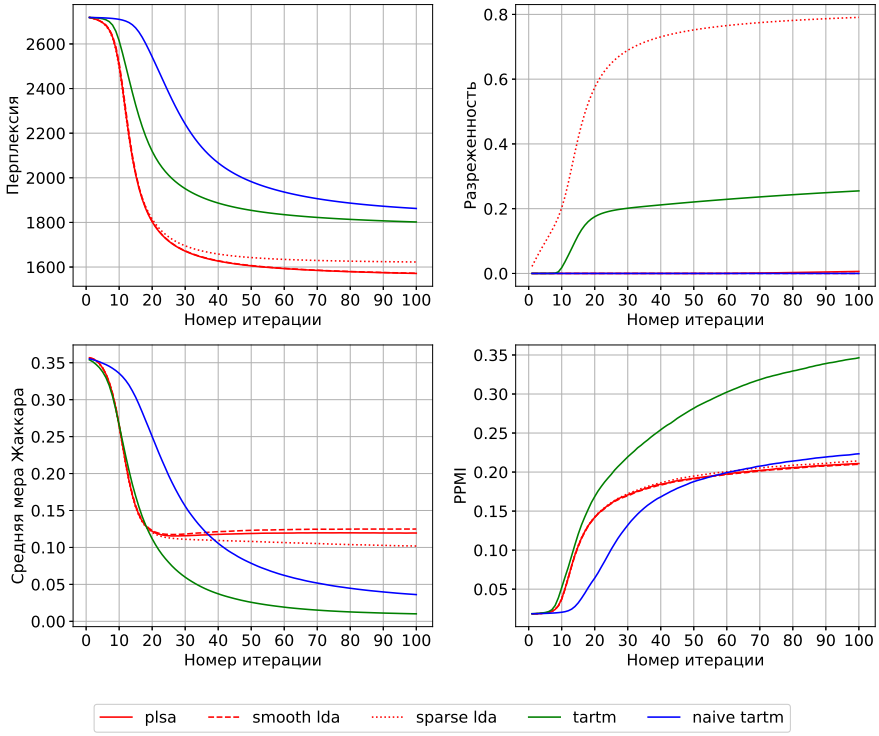

Рис. 3 — График зависимости трёх критериев качества тематических моделей (разреженности, средней различности тем по мере Жаккара, средней когерентности тем по PPMI) для пяти моделей (PLSA, LDA со сглаживанием, LDA с разреживанием, TARTM и «наивный» TARTM) на текстовой коллекции NIPS. Модель TARTM быстрее сходится, а по критериям различности и когерентности тем либо превосходит остальные модели, либо сравнима с ними.

В **заключении** приведены основные результаты работы, которые заключаются в следующем:

1. Теорема о достаточных условиях сходимости алгоритма ARTM.
2. Теорема о достаточных условиях единственности стохастического матричного разложения.
3. Модификация алгоритма ARTM, ускоряющая сходимость итерационного процесса.
4. Метод разреживания тематической модели, не увеличивающий перплексию получаемой модели.

Публикации автора по теме диссертации

В изданиях из списка ВАК РФ

Селезнев, *Н.* Автоматическое извлечение атрибутов водителя из логов мобильного приложения такси [Текст] / Н. Селезнев, И. Ирхин, В. Кантор // Труды МФТИ. — 2018. — Т. 10, № 3. — С. 5—15.

В изданиях, входящих в международную базу цитирования Scopus

Ирхин, *И.* Сходимость алгоритма аддитивной регуляризации тематических моделей [Текст] / И. Ирхин, К. Воронцов // Труды Института математики и механики УрО РАН. — 2020. — Т. 26, № 3. — С. 57—68.

Дербаносов, *Р. Ю.* Проблемы устойчивости и единственности стохастического матричного разложения [Текст] / Р. Ю. Дербаносов, И. Ирхин // Журнал вычислительной математики и математической физики. — 2020. — Т. 60, № 3. — С. 19—28. — (0,21 п. л., WoS).

Ирхин, *И.* Аддитивная регуляризация тематических моделей с быстрой векторизацией текста [Текст] / И. Ирхин, В. Булатов, К. Воронцов // Компьютерные исследования и моделирование. — 2020. — Т. 12, № 6.