КОМПЬЮТЕРНЫЕ ИССЛЕДОВАНИЯ
И МОДЕЛИРОВАНИЕ 2020 Т. 12 № 6 С. 1515–1528
DOI: 10.20537/2076-7633-2020-12-6-1515-1528

Ки&М

МОДЕЛИ ЭКОНОМИЧЕСКИХ И СОЦИАЛЬНЫХ СИСТЕМ

УДК: 004.852, 519.853

Аддитивная регуляризация тематических моделей с быстрой векторизацией текста

И. А. Ирхин[a], В. Г. Булатов[b], К. В. Воронцов[c]

Московский физико-технический институт,
141701, Московская область, г. Долгопрудный, Институтский переулок, д. 9

E-mail: [a] ilirhin@gmail.com, [b] bt.uytya@gmail.com, [c] k.v.vorontsov@phystech.edu

Получено 21.09.2020, после доработки — 01.10.2020.
Принято к публикации 05.10.2020.

Задача вероятностного тематического моделирования заключается в том, чтобы по заданной коллекции текстовых документов найти две матрицы: матрицу условных вероятностей тем в документах и матрицу условных вероятностей слов в темах. Каждый документ представляется в виде мультимножества слов, то есть предполагается, что для выявления тематики документа не важен порядок слов в нем, а важна только их частота. При таком предположении задача сводится к вычислению низкорангового неотрицательного матричного разложения, наилучшего по критерию максимума правдоподобия. Данная задача имеет в общем случае бесконечное множество решений, то есть является некорректно поставленной. Для регуляризации ее решения к логарифму правдоподобия добавляется взвешенная сумма оптимизационных критериев, с помощью которых формализуются дополнительные требования к модели. При моделировании больших текстовых коллекций хранение первой матрицы представляется нецелесообразным, поскольку ее размер пропорционален числу документов в коллекции. В то же время тематические векторные представления документов необходимы для решения многих задач текстовой аналитики — информационного поиска, кластеризации, классификации, суммаризации текстов. На практике тематический вектор вычисляется для каждого документа по необходимости, что может потребовать десятков итераций по всем словам документа. В данной работе предлагается способ быстрого вычисления тематического вектора для произвольного текста, требующий лишь одной итерации, то есть однократного прохода по всем словам документа. Для этого в модель вводится дополнительное ограничение в виде уравнения, позволяющего вычислять первую матрицу через вторую за линейное время. Хотя формально данное ограничение не является оптимизационным критерием, фактически оно выполняет роль регуляризатора и может применяться в сочетании с другими критериями в рамках теории аддитивной регуляризации тематических моделей ARTM. Эксперименты на трех свободно доступных текстовых коллекциях показали, что предложенный метод улучшает качество модели по пяти оценкам качества, характеризующим разреженность, различность, информативность и когерентность тем. Для проведения экспериментов использовались библиотеки с открытым кодом BigARTM и TopicNet.

Ключевые слова: автоматическая обработка текстов, обучение без учителя, тематическое моделирование, аддитивная регуляризация тематических моделей, EM-алгоритм, PLSA, LDA, ARTM, BigARTM, TopicNet

Работа выполнена в рамках проекта «Средства интеллектуального анализа больших массивов текстов», по Программе ЦК НТИ «Центр хранения и анализа больших данных», поддерживаемого Министерством науки и высшего образования Российской Федерации по договору от 15.08.2019 № 7/1251/2019. Работа также частично поддержана РФФИ, проект 20-07-00936.

COMPUTER RESEARCH AND MODELING
2020 VOL. 12 NO. 6 P. 1515–1528
DOI: 10.20537/2076-7633-2020-12-6-1515-1528

Ки&М

MODELS OF ECONOMIC AND SOCIAL SYSTEMS

UDC: 004.852, 519.853

Additive regularizarion of topic models with fast text vectorizartion

I. A. Irkhin[a], V. G. Bulatov[b], K. V. Vorontsov[c]

Moscow Institute of Physics and Technology,
9 Institutskiy per., Dolgoprudny, Moscow oblast, Dolgoprudny, 141701, Russia

E-mail: [a] ilirhin@gmail.com, [b] bt.uytya@gmail.com, [c] k.v.vorontsov@phystech.edu

Received 21.09.2020, after completion — 01.10.2020.
Accepted for publication 05.10.2020.

The probabilistic topic model of a text document collection finds two matrices: a matrix of conditional probabilities of topics in documents and a matrix of conditional probabilities of words in topics. Each document is represented by a multiset of words also called the "bag of words", thus assuming that the order of words is not important for revealing the latent topics of the document. Under this assumption, the problem is reduced to a low-rank non-negative matrix factorization governed by likelihood maximization. In general, this problem is ill-posed having an infinite set of solutions. In order to regularize the solution, a weighted sum of optimization criteria is added to the log-likelihood. When modeling large text collections, storing the first matrix seems to be impractical, since its size is proportional to the number of documents in the collection. At the same time, the topical vector representation (embedding) of documents is necessary for solving many text analysis tasks, such as information retrieval, clustering, classification, and summarization of texts. In practice, the topical embedding is calculated for a document "on-the-fly", which may require dozens of iterations over all the words of the document. In this paper, we propose a way to calculate a topical embedding quickly, by one pass over document words. For this, an additional constraint is introduced into the model in the form of an equation, which calculates the first matrix from the second one in linear time. Although formally this constraint is not an optimization criterion, in fact it plays the role of a regularizer and can be used in combination with other regularizers within the additive regularization framework ARTM. Experiments on three text collections have shown that the proposed method improves the model in terms of sparseness, difference, logLift and coherence measures of topic quality. The open source libraries BigARTM and TopicNet were used for the experiments.

Keywords: natural language processing, unsupervised learning, topic modeling, additive regularization of topic model, EM-algorithm, PLSA, LDA, ARTM, BigARTM, TopicNet

Citation: *Computer Research and Modeling*, 2020, vol. 12, no. 6, pp. 1515–1528 (Russian).

The work was carried out within the project "Intelligent data analysis tools for large text collections" under the NTI Program "Center for big data storage and analysis" supported by the Ministry of Science and Higher Education of the Russian Federation under the agreement of 15.08.2019 No. 7/1251/2019. Also the work is partially supported by RFBR, project 20-07-00936

Введение

Тематическое моделирование — одно из современных направлений обработки естественного языка (natural language processing, NLP). Вероятностная тематическая модель (probabilistic topic model, PTM) коллекции текстовых документов описывает каждый документ d дискретным распределением вероятностей тем $\theta_{td} = p(t|d)$, каждую тему t — дискретным распределением вероятностей слов $\phi_{wt} = p(w|t)$. Тематическая модель выявляет кластерную структуру текстовой коллекции, однако в отличие от обычной «жесткой» кластеризации каждый документ относится не к одному кластеру, а к нескольким кластерам-темам. Вектор условных вероятностей тем в документе (тематический вектор) может быть использован в качестве числового признакового описания документа для решения различных задач текстовой аналитики, в том числе для поиска тематически близких документов [Yanina et al., 2018; Ianina, Vorontsov, 2019], классификации текстов [Rubin et al., 2012], восстановления регрессионных зависимостей на текстах [Sokolov, Bogolubsky, 2015], тематической сегментации текстов [Riedl, Biemann, 2012; Skachkov, Vorontsov, 2018], суммаризации текстов [Litvak et al., 2015]. В этих и многих других задачах преимуществом тематического моделирования является возможность описать каждую тему — компоненту тематического вектора — словами или фразами естественного языка. Это свойство тематических векторных представлений текста, называемое интерпретируемостью, важно для многих приложений [Boyd et al., 2017].

Доминирующим подходом к тематическому моделированию в настоящее время считается байесовское обучение. В байесовском подходе любые требования к модели формализуются через априорные распределения параметров модели и структуру взаимосвязей между скрытыми переменными. Это приводит к тому, что для любой новой модели необходимо заново проделывать байесовский вывод, затем разработку, реализацию и тестирование алгоритма [Boyd et al., 2017]. При этом не существует простого способа комбинирования нескольких моделей, для которых байесовский вывод и реализация алгоритма уже были проделаны ранее. Теория аддитивной регуляризации тематических моделей ARTM решает эту проблему, отказываясь от использования байесовского вывода [Воронцов, 2014; Vorontsov, Potapenko, 2015]. В ARTM любые требования к модели формализуются через оптимизационные критерии — регуляризаторы. Если требований несколько, то в постановку оптимизационной задачи вводится взвешенная сумма регуляризаторов. Байесовские тематические модели, как правило, удается переформулировать в терминах регуляризации, при этом существенно сокращается объем необходимых математических выкладок [Kochedykov et al., 2017]. Для оценивания параметров модели с произвольным набором регуляризаторов используется один и тот же итерационный процесс, называемый регуляризованным EM-алгоритмом. Появляется 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` [Vorontsov et al., 2015; Frei, Apishev, 2016] и `TopicNet` [Bulatov et al., 2020].

В обоих подходах, байесовском и регуляризационном, основными параметрами тематической модели являются две матрицы: матрица условных вероятностей тем в документах $\Theta = (\theta_{td})$ и матрица условных вероятностей слов в темах $\Phi = (\phi_{wt})$. Вычисление тематического вектора документа θ_d обычно требует десятков итераций. В данной работе предлагается простой способ вычисления матрицы Θ по матрице Φ за одну итерацию. Функциональная связь $\Theta = f(\Phi)$ позволяет отказаться от хранения матрицы Θ и вычислять тематические векторы для любых документов или текстовых фрагментов за время, линейное по числу слов в тексте. Данная функциональная связь не является оптимизационным критерием регуляризации, однако она приводит к модификации EM-алгоритма, аналогичной регуляризаторам. В рамках подхода ARTM она может использоваться наравне и в сочетании с другими регуляризаторами.

Матрица Θ не хранится также в онлайновых алгоритмах для моделей PLSA [Bassiou, Kotropoulos, 2014], LDA [Hoffman et al., 2010] и широкого класса регуляризованных моделей ARTM [Vorontsov et al., 2015; Frei, Apishev, 2016]. Онлайновые версии EM-алгоритма обрабатывают большую коллекцию текстов за время, линейное по числу слов в коллекции. Однако обработка каждого документа в этих алгоритмах требует многих итераций по всем словам документа (в отличие от предлагаемого метода, который обходится одной итерацией).

Отказ от хранения матрицы Θ приводит к сокращению размерности тематической модели и уменьшению переобучения. В моделях с двумя матрицами недостаточное качество матрицы Φ теоретически может быть скомпенсировано итерационным процессом подгонки каждого столбца θ_d матрицы Θ под конкретный документ d. Когда матрица Θ непосредственно зависит от матрицы Φ, такая подгонка становится невозможной. Кроме того, размер матрицы Θ линейно зависит от числа документов в коллекции, тогда как размер словаря увеличивается по сублинейному степенному закону Хипса [Egghe, 2007]. Рост словаря может быть ограничен и принудительно, путем отбрасывания наименее частотных слов. Таким образом, не только размерность модели уменьшается, но и сокращается темп ее роста при расширении коллекции. Наконец, именно матрица Φ обычно используется для оценивания интерпретируемости моделей [Chang et al., 2009; Newman et al., 2010b; Röder et al., 2015], что также является косвенным подтверждением избыточности, вторичности матрицы Θ по отношению к основной матрице параметров тематической модели Φ.

Эксперименты, описанные в последнем разделе данной работы, подтверждают, что предлагаемый способ быстрой векторизации текста улучшает интерпретируемость, различность и разреженность тем одновременно по нескольким критериям.

Аддитивная регуляризация тематических моделей

Пусть заданы три конечных множества: D — коллекция текстовых документов, W — словарь термов, T — множество тем. *Термами* могут быть как слова, так и нормальные формы слов, словосочетания или термины предметной области (в зависимости от того, какие виды предварительной обработки текстов были выполнены). Каждый документ $d \in D$ представляет собой последовательность n_d термов $(w_1, \ldots, w_{n_d})$ из словаря W. Гипотеза «мешка слов» предполагает, что порядок слов в документе не важен, поэтому документ более компактно представляется мультимножеством $d \subset W$, в котором каждый терм $w \in d$ встречается n_{dw} раз.

Тематическая модель описывает условные вероятности $p(w|d)$ появления термов w в документах d через вероятности термов в темах $\phi_{wt} = p(w|t)$ и 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,

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

Для оценивания параметров тематической модели $\Phi = (\phi_{wt})_{W\times T}$ и $\Theta = (\theta_{td})_{T\times D}$ по коллекции D максимизируется логарифм правдоподобия с регуляризатором [Воронцов, 2014; Vorontsov, Potapenko, 2015]:

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

при ограничениях неотрицательности и нормировки на столбцы матриц Φ и Θ

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

Таким образом, построение вероятностной тематической модели текстовой коллекции — это задача приближенного низкорангового стохастического матричного разложения. Данная задача является некорректно поставленной по Адамару, поскольку множество ее решений в общем случае бесконечно. Регуляризатор R позволяет выбрать из всего множества решений более приемлемые. Если критериев несколько, то R задается в виде взвешенной комбинации регуляризаторов.

Наиболее известные тематические модели PLSA и LDA являются частными случаями регуляризации. В вероятностном латентном семантическом анализе (probabilistic latent semantic analysis) PLSA [Hofmann, 1999] регуляризатор нулевой, $R(\Phi, \Theta) = 0$. В латентном размещении Дирихле (latent Dirichlet allocation) LDA [Blei et al., 2003] 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 неотрицательными гиперпараметрами β_w, α_t, которые на практике обычно фиксируются. При $\beta_w < 1$ и $\alpha_t < 1$ данный регуляризатор приводит к разреживанию распределений ϕ_{wt} и θ_{td}. Обзор полезных регуляризаторов можно найти в [Kochedykov et al., 2017].

Обозначим через $\underset{i\in I}{\text{norm}}\, x_i = \dfrac{\max\{x_i, 0\}}{\sum\limits_{j\in I}\max\{x_j, 0\}}$ операцию нормирования вектора $(x_i\colon i \in I)$.

Теорема 1 ([Воронцов, 2014; Vorontsov, Potapenko, 2015]). *Решение* Φ, Θ *задачи* (2)–(3) *удовлетворяет следующей системе уравнений относительно переменных* ϕ_{wt}, θ_{td} *и* p_{tdw}*:*

$$p_{tdw} = \underset{t\in T}{\text{norm}}(\phi_{wt}\theta_{td}), \tag{4}$$

$$\phi_{wt} = \underset{w\in W}{\text{norm}}\Big(\sum_{d\in D} n_{dw}p_{tdw} + \phi_{wt}\frac{\partial R}{\partial\phi_{wt}}\Big), \tag{5}$$

$$\theta_{td} = \underset{t\in T}{\text{norm}}\Big(\sum_{w\in d} n_{dw}p_{tdw} + \theta_{td}\frac{\partial R}{\partial\theta_{td}}\Big). \tag{6}$$

Вероятностный смысл переменных p_{tdw} следует из формулы Байеса и гипотезы условной независимости — это распределение тем для терма w в документе d:

$$p_{tdw} = \frac{\phi_{wt}\theta_{td}}{\sum\limits_{s\in T}\phi_{ws}\theta_{sd}} = \frac{p(w|t)p(t|d)}{p(w|d)} = \frac{p(t, w|d)}{p(w|d)} = p(t|d, w).$$

Распределение вида $p(t|x)$ будем называть *тематическим вектором* объекта x. В частности, можно говорить о тематических векторах документа $p(t|d)$, терма $p(t|w)$, терма в документе $p(t|d, w)$.

Решение системы (4)–(6) методом простых итераций приводит к регуляризованному EM-алгоритму, который состоит в чередовании E-шага (4) и M-шага (5)–(6) до сходимости. В реализации `BigARTM` [Vorontsov et al., 2015] начальное приближение задается равномерным для $\theta_{td}^0 = \dfrac{1}{|T|}$ и случайным для $\phi_{wt}^0 = \underset{w}{\text{norm}}(\text{rand}[0, 1])$. Одна итерация EM-алгоритма требует однократного прохода по всем термам всех документов.

Чтобы добавить новый регуляризатор в алгоритм, достаточно добавить его частные производные по параметрам модели в формулы M-шага. Таким образом, ARTM — это не одна частная модель или метод, а общий оптимизационный подход к построению и комбинированию широкого класса тематических моделей.

EM-алгоритм с быстрой векторизацией документов

Потребуем, чтобы тематический вектор произвольного документа d совпадал с результатом первой итерации EM-алгоритма без регуляризации при равномерном начальном приближении $\theta_{td}^0 = \dfrac{1}{|T|}$:

$$\theta_{td}(\Phi) = \underset{t\in T}{\operatorname{norm}}\Big(\sum_{w\in W} n_{dw}p_{tdw}\Big) = \sum_{w\in d}\frac{n_{dw}}{n_d}\frac{\phi_{wt}\theta_{td}^0}{\sum_s \phi_{ws}\theta_{sd}^0} = \sum_{w\in d} p_{wd}\frac{\phi_{wt}}{\sum_s \phi_{ws}}, \tag{7}$$

где $p_{wd} = \dfrac{n_{dw}}{n_d} = \hat{p}(w|d)$ — частотная оценка условной вероятности терма в документе.

Уравнение (7) позволяет исключить переменные θ_{td} из постановки задачи. Для этого нам будет удобно перейти к другой форме записи М-шага.

Лемма 1. *При фиксированных значениях переменных p_{tdw} уравнения М-шага* (5)–(6) *являются необходимыми условиями экстремума оптимизационной задачи*

$$Q(\Phi,\Theta) = \sum_{d\in D}\sum_{w\in W}\sum_{t\in T} n_{dw}p_{tdw}\ln(\phi_{wt}\theta_{td}) + R(\Phi,\Theta) \;\to\; \max_{\Phi,\Theta}.$$

Доказательство следует из условий Каруша – Куна – Таккера для задачи максимизации $Q(\Phi,\Theta)$ при ограничениях (3), аналогично доказательству теоремы 1 (см. [Vorontsov, Potapenko, 2015]).

В теории обобщенного EM-алгоритма (generalized EM, GEM) [Dempster et al., 1977] функционал Q является нижней оценкой регуляризованного логарифма правдоподобия (2), поэтому оптимальное значение Q монотонно увеличивается на каждой EM-итерации.

Рассмотрим сначала общий случай функциональной зависимости $\theta_{td}(\Phi)$.

Теорема 2. *Пусть функции $\theta_{td}(\Phi)$ и $R(\Phi,\Theta)$ непрерывно дифференцируемы. Тогда точка Φ локального экстремума задачи* (2)–(3) *с дополнительными ограничениями-равенствами $\theta_{td} = \theta_{td}(\Phi)$ удовлетворяет системе уравнений со вспомогательными переменными $p_{tdw} = p(t|d,w)$, n_{td} и n_{wt}:*

$$p_{tdw} = \underset{t\in T}{\operatorname{norm}}(\phi_{wt}\theta_{td}), \tag{8}$$

$$n_{td} = \sum_{w\in d} n_{dw}p_{tdw}, \tag{9}$$

$$n_{wt} = \sum_{d\in D} n_{dw}p_{tdw}, \tag{10}$$

$$\phi_{wt} = \underset{w\in W}{\operatorname{norm}}\Big(n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}} + \phi_{wt}\sum_{d\in D}\sum_{s\in T}\Big(\frac{n_{sd}}{\theta_{sd}} + \frac{\partial R}{\partial \theta_{sd}}\Big)\frac{\partial \theta_{sd}}{\partial \phi_{wt}}\Big). \tag{11}$$

Доказательство. Введем функционал $\tilde{Q}(\Phi)$, подставив выражение $\theta_{dt}(\Phi)$ в $Q(\Phi,\Theta)$:

$$\tilde{Q}(\Phi) = \sum_{d\in D}\sum_{u\in W}\sum_{s\in T} n_{du}p_{sdu}(\ln\phi_{us} + \ln\theta_{sd}(\Phi)) + R(\Phi,\Theta(\Phi)) \;\to\; \max_{\Phi}.$$

Запишем частные производные $\tilde{Q}$ по параметрам ϕ_{wt}, выделяя в формулах выражения n_{wt} и n_{sd} согласно (10) и (9) соответственно:

$$\phi_{wt}\frac{\partial\tilde{Q}}{\partial\phi_{wt}} = \sum_{d\in D} n_{dw}p_{tdw} + \sum_{d,s,u} n_{du}p_{sdu}\frac{\phi_{wt}}{\theta_{sd}}\frac{\partial\theta_{sd}}{\partial\phi_{wt}} + \phi_{wt}\frac{\partial R}{\partial\phi_{wt}} + \phi_{wt}\sum_{d,s}\frac{\partial R}{\partial\theta_{sd}}\frac{\partial\theta_{sd}}{\partial\phi_{wt}} =$$
$$= n_{wt} + \phi_{wt}\frac{\partial R}{\partial\phi_{wt}} + \phi_{wt}\sum_{d,s}\Big(\frac{n_{sd}}{\theta_{sd}} + \frac{\partial R}{\partial\theta_{sd}}\Big)\frac{\partial\theta_{sd}}{\partial\phi_{wt}}.$$

Записывая условия Каруша–Куна–Таккера, по аналогии с [Vorontsov, Potapenko, 2015]

$$\phi_{wt} = \underset{w\in W}{\operatorname{norm}}\left(\phi_{wt}\frac{\partial\tilde{Q}}{\partial\phi_{wt}}\right),$$

получаем искомое уравнение (11). Теорема доказана. □

Таким образом, модификация EM-алгоритма коснулась только формул M-шага, причем аддитивная поправка к частотным оценкам условных вероятностей $p(w|t)$ имеет вид, аналогичный регуляризационным поправкам.

Теперь вернемся к частному случаю (7). Чтобы применить к нему теорему 2, найдем частную производную:

$$\frac{\partial\theta_{sd}}{\partial\phi_{wt}} = \frac{\partial}{\partial\phi_{wt}}\Big(\frac{p_{wd}\phi_{ws}}{\sum_v\phi_{wv}}\Big) = p_{wd}\frac{\delta_{st}\sum_v\phi_{wv} - \phi_{ws}}{(\sum_v\phi_{wv})^2} = p_{wd}h_w(\delta_{st} - \phi_{ws}h_w),$$

где $\delta_{st} = [s{=}t]$ — символ Кронекера, $h_w = (\sum_t\phi_{wt})^{-1}$. Подставим это выражение в (11) и перепишем уравнения в порядке, удобном для последовательного проведения вычислений в методе простых итераций:

$$h_w = (\textstyle\sum_t\phi_{wt})^{-1},$$
$$\theta_{td} = \sum_{w\in d} p_{wd}\phi_{wt}h_w,$$
$$p_{tdw} = \underset{t\in T}{\operatorname{norm}}(\phi_{wt}\theta_{td}),$$
$$c_{td} = \frac{1}{\theta_{td}}\sum_{w\in d} n_{dw}p_{tdw} + \frac{\partial R}{\partial\theta_{td}},$$
$$\gamma_{dw} = \sum_{t\in T}\phi_{wt}c_{td},$$
$$p'_{tdw} = p_{tdw} + n_d^{-1}\phi_{wt}h_w(c_{td} - h_w\gamma_{dw}),$$
$$\phi_{wt} = \underset{w\in W}{\operatorname{norm}}\Big(\sum_{d\in D} n_{dw}p'_{tdw} + \phi_{wt}\frac{\partial R}{\partial\phi_{wt}}\Big).$$

Вычисление переменных θ_{td}, p_{tdw}, c_{td}, γ_{dw} и p'_{tdw} образует E-шаг и занимает $O(n_d|T|)$ операций для каждого документа d, как и в обычном EM-алгоритме. Все эти переменные, относящиеся к конкретному документу d, можно удалять из памяти по окончании его обработки. Вычисление переменных ϕ_{wt} происходит по обычной формуле, за исключением того, что вместо условных вероятностей p_{tdw} подставляются переменные p'_{tdw}. Таким образом, модификация EM-алгоритма не приводит к существенному увеличению ни времени его работы, ни расхода памяти.

Обработка каждого документа делается за два прохода. На первом проходе вычисляются только переменные θ_{td}. На втором проходе вычисляются все остальные переменные, используемые в конечном итоге для вычисления p'_{tdw} и обновления ϕ_{wt}. Если требуется только найти

тематический вектор документа θ_d, не обновляя матрицу Φ, то второй проход делать не нужно. Переменные h_w также можно не вычислять каждый раз, формируя их после очередного обновления матрицы Φ. Таким образом, рассмотренная модификация EM-алгоритма позволяет тематизировать новые документы максимально быстро.

Оценивание качества тематических моделей

Для оценивания тематических моделей в экспериментах на текстовых коллекциях будем использовались следующие критерии качества.

Разреженность матрицы Φ определяется как доля нулевых элементов в матрице. Высокая разреженность свидетельствует о том, что в каждой теме выделилось семантическое ядро — множество термов, имеющих ненулевую вероятность. Косвенно это говорит об успешной кластеризации терминов по темам [Vorontsov, Potapenko, 2014].

Когерентность темы показывает, насколько часто ключевые слова темы сочетаются в одних и тех же контекстах. В качестве ключевых слов темы t обычно берут k термов с наибольшей вероятностью $p(w|t)$. В наших экспериментах $k = 30$. Мерой сочетаемости термов w и v является положительная поточечная взаимная информация (positive point-wise mutual information, PPMI):

$$\text{PPMI}(w, v) = \max\left(0, \log \frac{|D|N(w, v)}{N(w)N(v)}\right),$$

где $N(w, v)$ — число документов, в которых встречаются оба терма, w и v; $N(w)$ — число документов, содержащих терм w. Когерентность модели определяется как средний $\text{PPMI}(w, v)$ по всем темам и всем парам ключевых термов в каждой теме. Известно, что когерентность является адекватной мерой интерпретируемости тематической модели, поскольку лучше других критериев коррелирует с экспертными оценками интерпретируемости [Newman et al., 2010a; Newman et al., 2010b; Lau et al., 2014; Röder et al., 2015]. Чем выше когерентность, тем лучше.

Сходство тем по множествам ключевых слов. Для каждой темы t выбирается подмножество $W_t \subset W$ из k термов с наибольшими значениями вероятности $p(w|t)$. В наших экспериментах $k = 100$. Среднее попарное сходство тем в модели определяется как мера Жаккара между множествами W_t и W_s, усредненная по всем парам тем t, s:

$$\frac{1}{|T|(|T| - 1)} \sum_{s \neq t} \frac{|W_t \cap W_s|}{|W_t \cup W_s|}.$$

Если мера Жаккара равна 1, то две темы являются дубликатами. Чем меньше среднее попарное сходство тем, тем лучше.

Среднее расстояние до ближайшей темы. Это другой способ оценивания различности тем, в котором учитываются все распределения $p(w|t)$, а не только ключевые термы темы. Для каждой темы определяется ближайшая к ней тема по косинусному расстоянию, эти расстояния усредняются по всем темам. Косинусное расстояние принимает значения от 0 до 1. Расстояние 1 соответствует ортогональным темам, не имеющим общей лексики. Чем больше среднее расстояние до ближайшей темы, тем лучше.

Средний logLift. Критерий logLift показывает, насколько терм w важен для темы t:

$$\log \text{Lift}(w, t) = \log \frac{p(w|t)}{p(w)}.$$

Он был предложен в [Taddy, 2012] в качестве критерия сортировки термов при визуализации темы в пользовательском интерфейсе. В недавней работе [Fan et al., 2019] было предложено усреднять logLift по $k = 30$ ключевым словам в каждой теме, затем по всем темам. Было показано, что

вычисляемый таким способом средний logLift связан с долей неинформативных слов в темах, а также что он имеет существенную корреляцию с экспертными оценками качества тем. Чем больше средний logLift, тем лучше.

Эксперименты

Эксперименты проводились на трех общедоступных текстовых коллекциях: 20 newsgroups (были отобраны 8 новостных групп: auto, motorcycles, baseball, hockey, crypt, electronics, med, space), статьи конференции NIPS с 1987 по 2015 г. и Twitter Sentiment 140. Число тем было выбрано $|T| = 25$ для коллекции 20 newsgroups, $|T| = 50$ для коллекций NIPS и Twitter.

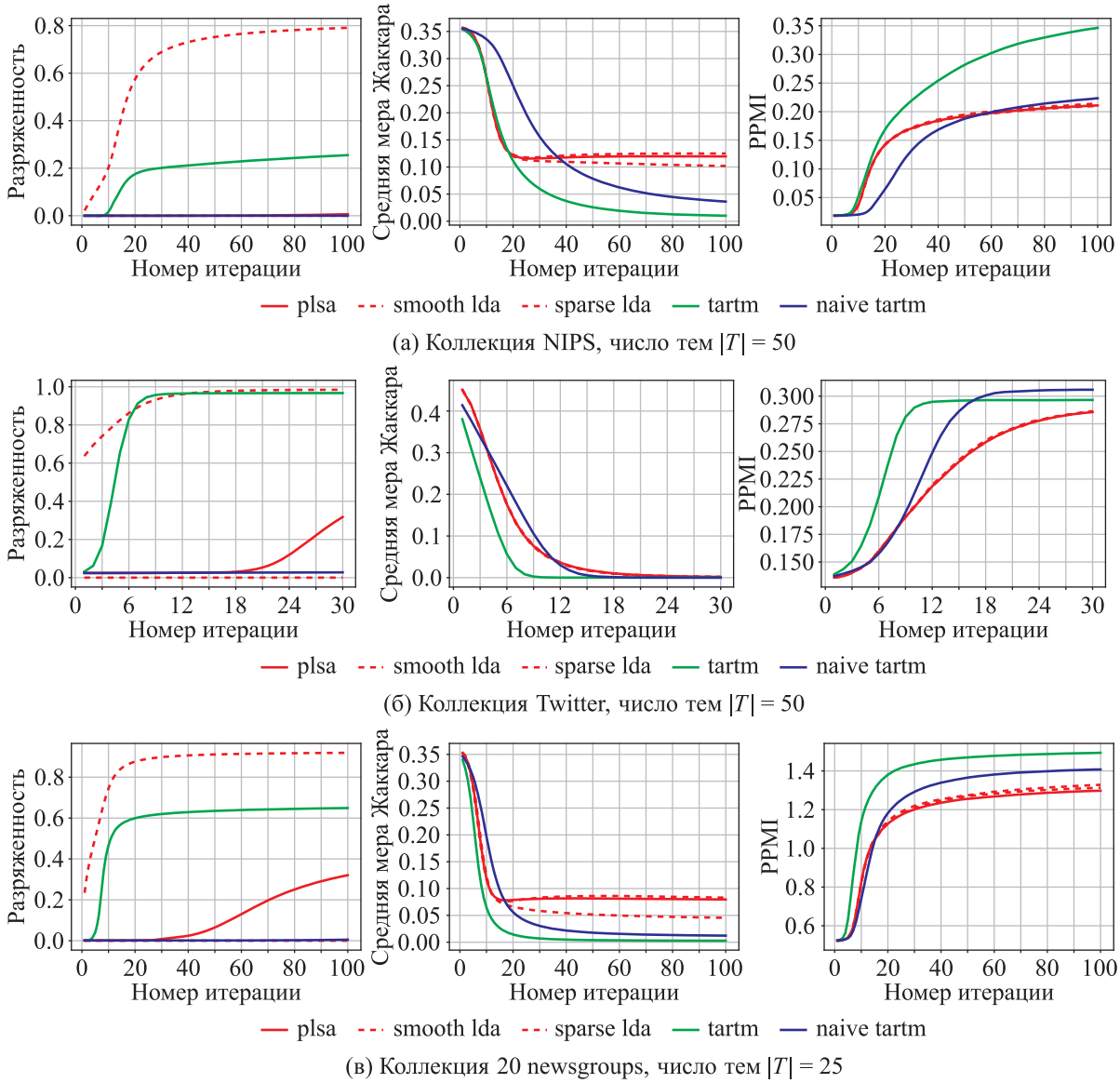
Сравнивались следующие тематические модели: PLSA, два варианта LDA со сглаживающим и разреживающим априорным распределением Дирихле (`smooth LDA` и `sparse LDA` соответственно) и две модели с быстрым вычислением тематических векторов документов: `TARTM` — алгоритм из теоремы 2 (thetaless ARTM) и «наивный» TARTM (`naive TARTM`). Последний представляет собой обычный EM-алгоритм, в котором вектор θ_d вычисляется за одну итерацию из начального приближения $\theta_{td} = \dfrac{1}{|T|}$ по формуле (6). Сравнение с «наивным» TARTM необходимо для того, чтобы проверить, не решается ли поставленная задача быстрой векторизации документов простым, очевидным способом.

Для проведения численных экспериментов был реализован отдельный модуль на языке Python, позволяющий сравнивать различные варианты реализации[1]. Кроме того, мы использовали возможность добавления пользовательского регуляризатора в проекте с открытым кодом `TopicNet`. `TopicNet` — это надстройка над библиотекой `BigARTM`, предоставляющая дополнительную функциональность для построения и анализа регуляризованных тематических моделей [Bulatov et al., 2020]. Встраивание еще одного регуляризатора как в `BigARTM`, так и в `TopicNet` сводится к реализации функции, вычисляющей еще одно слагаемое в формуле M-шага. Заметим, что оно не обязано иметь вид $\phi_{wt}\dfrac{\partial R}{\partial \phi_{wt}}$, то есть выражаться через частные производные некоторой функции R. В нашем случае это действительно важное замечание, поскольку модифицированный M-шаг не соответствует никакому критерию регуляризации R. Чтобы каждый столбец θ_d вычислялся по Ф за одну итерацию EM-алгоритма, то есть по формуле (7), мы устанавливали параметр `num_document_passes = 1`, предоставляемый библиотекой `BigARTM`.

Цель первого эксперимента состояла в сравнении сходимости пяти моделей по нескольким критериям. Результаты показаны на рис. 1. Модели `TARTM` демонстрируют наилучший результат по когерентности и различности тем (средней мере Жаккара). Высокую различность можно объяснить тем, что `TARTM` очищает темы от общеупотребительных слов (в отличие от LDA). В таблице 1 это показано на примере трех тем. По критерию разреженности `TARTM` уступает разреживающей модели LDA, однако он и не стремится разреживать модель в явном виде (для этого можно было бы ввести в `TARTM` регуляризатор разреживания). Тем не менее `TARTM` существенно превосходит остальные модели, в которые также не ставится задача разреживания.

Целью второго эксперимента было сравнение `TARTM` с другими регуляризованными моделями. Обычным рекомендуемым набором регуляризаторов является комбинация сглаживания фоновых тем, разреживания предметных тем и декоррелирования [Vorontsov, Potapenko, 2015]. Для сравнения использовалась реализация в `TopicNet`. Модели сравнивались не на каждой итерации, а только после всех итераций; число тем полагалось равным 20. Результаты приведены

[1] Исходный код экспериментов доступен по адресу `github.com/ilirhin/python_artm/tree/master/pyartm_experiments/pyartm_experiments/thetaless` и `github.com/machine-intelligence-laboratory/TopicNet/blob/master/topicnet/demos/Topic-Thetaless-Regularizer.ipynb`



(а) Коллекция NIPS, число тем $|T| = 50$

(б) Коллекция Twitter, число тем $|T| = 50$

(в) Коллекция 20 newsgroups, число тем $|T| = 25$

Рис. 1. Графики зависимости трех 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) по трем текстовым коллекциям (NIPS, Twitter, 20 newsgroups). Модель TARTM быстрее сходится, а по критериям различности и когерентности тем либо превосходит остальные модели, либо сравнима с ними

в таблице 2. Они показывают, что предложенная модификация М-шага для быстрой векторизации текстов может эффективно комбинироваться с другими регуляризаторами и улучшать качество модели.

Полученные результаты можно интерпретировать следующим образом. В обычных тематических моделях обе матрицы — Φ и Θ — оптимизируются для предсказания слов в документах, тогда как `TARTM` использует для этого только матрицу Φ. Поэтому обычные модели имеют возможность скомпенсировать недостатки матрицы Φ путем переобучения, то есть с помощью избыточно точной подгонки столбцов матрицы Θ под конкретные документы, в то время как `TARTM` имеет возможность исправлять эти недостатки, только улучшая основную матрицу параметров модели Φ.

Таблица 1. Примеры наиболее частотных слов в темах на коллекции 20 newsgroups. Слова общей лексики выделены жирным шрифтом. `TARTM` выделяет слова общей лексики в отдельные темы (в отличие от модели LDA)

TARTM	LDA
game team player play season hockey hit league fan baseball **last** run watch throw pitcher ball stat year sport score	game year team player get **last good** baseball win play **go** season hit fan **think** time **make well say** league
car bike buy engine sell speed drive price mile road ride owner dealer drive model driver motorcycle tire detector brake	car bike **get** engine buy new **also** drive mile **make** speed look tire **well** dealer brake wheel **go good** road
period st series vs playoff pt shot king canada ranger lead cup toronto play wing pittsburgh buffalo blue chicago round	period gm vs pt st chicago power pp april shot play buffalo pittsburgh islander flame series lead **first** scorer cup

Таблица 2. Эксперимент с реализацией `TopicNet`. Сравнение моделей по пяти критериям: sparsity — разреженность матрицы Φ, PPMI — средняя когерентность, Jaccard — среднее сходство тем по мере Жаккара, minDist — среднее расстояние до ближайшей темы, logLift — средний 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

Модель	sparsity	PPMI	Jaccard	minDist	LogLift
sparse LDA	0.896	1.570	0.044	0.587	0.503
smooth LDA	0	1.509	0.043	0.632	0.479
PLSA	0.869	1.517	0.050	0.586	0.459
ARTM + *Reg*	0.898	1.710	0.027	0.661	0.590
TARTM	0.893	1.716	0.007	0.895	0.952
TARTM + *Reg*	**0.929**	**1.788**	**0.003**	**0.953**	**1.020**

Заключение

Низкоранговые матричные разложения применяются в самых разных предметных областях для сокращения размерности данных, то есть для преобразования векторных представлений объектов высокой размерности в векторы существенно более низкой размерности при минимальных потерях точности представления. В случае вероятностного тематического моделирования текстовые документы исходно представляются дискретными распределениями на множестве слов; размерность таких векторных представлений обычно составляет десятки или сотни тысяч. Стохастическое матричное разложение позволяет представлять документы дискретными распределениями на множестве тем; их размерность может составлять десятки или сотни, реже — тысячи; она задается самим исследователем из соображений вычислительной эффективности либо оптимизируется по внешним критериям, таким как качество классификации или поиска текстовых документов.

Во многих приложениях низкоранговых матричных разложений к анализу больших данных размерности исходной матрицы имеют различный темп роста при увеличении объема данных. В случае моделирования больших текстовых коллекций число слов в словаре растет существенно медленнее по сравнению с числом документов в коллекции (обычно по сублинейному степенному закону Хипса). Соответственно, размеры двух матриц разложения также увеличиваются с разной скоростью. Для сокращения размерности модели становится выгодно избавиться

от той матрицы, которая растет быстрее (в нашем случае — Θ), научившись вычислять ее по второй матрице (Φ).

В данной работе предложена модификация оптимизационной задачи стохастического матричного разложения путем введения явного ограничения-равенства $\Theta = f(\Phi)$. Теорема 2 дает удобное обобщение EM-алгоритма, в котором данное ограничение принимает вид регуляризатора и может использоваться в сочетании с другими регуляризаторами в рамках теории аддитивной регуляризации тематических моделей (ARTM). Эксперименты подтверждают, что вместе они улучшают качество тематической модели по совокупности пяти критериев, характеризующих различность, разреженность и интерпретируемость тем.

В данной работе использован только один частный вид ограничения f, мотивированный формулой M-шага для одной итерации EM-алгоритма. Возможность применения других видов ограничения f пока остается открытым вопросом, равно как и возможность применения данной техники в других прикладных областях за пределами тематического моделирования. Открытым вопросом остается также адаптация предложенного подхода к мультимодальным тематическим моделям [Vorontsov et al., 2015], в которых имеется несколько матриц Φ.

Список литературы (References)

Воронцов К. В. Аддитивная регуляризация тематических моделей коллекций текстовых документов // *Доклады РАН*. — 2014. — Т. 455, № 3. — С. 268–271.
Vorontsov K. V. Additivnaya regulyarizaciya tematicheskih modelej kollekcij tekstovyh dokumentov [Additive regularization for topic models of text collections] // Doklady RAN. — 2014. — Vol. 455, No. 3. — P. 268–271 (in Russian).

Bassiou N., Kotropoulos C. Online PLSA: Batch updating techniques including out-of-vocabulary words // *Neural Networks and Learning Systems, IEEE Transactions on*. — Nov 2014. — Vol. 25, No. 11. — P. 1953–1966.

Blei D. M., Ng A. Y., Jordan M. I. Latent Dirichlet allocation // *Journal of Machine Learning Research*. — 2003. — Vol. 3. — P. 993–1022.

Boyd-Graber J., Hu Y., Mimno D. Applications of Topic Models. Foundations and Trends(r) in Information Retrieval Series. — now Publishers Incorporated, 2017.

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. — 2020. — P. 6745–6752.

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. — P. 288–296.

Dempster A. P., Laird N. M., Rubin D. B. Maximum likelihood from incomplete data via the EM algorithm // *J. of the Royal Statistical Society, Series B*. — 1977. — No. 34. — P. 1–38.

Egghe L. Untangling Herdan's law and Heaps' law: Mathematical and informetric arguments // *Journal of the American Society for Information Science and Technology*. — 2007. — Vol. 58, No. 5. — P. 702–709.

Fan A., Doshi-Velez F., Miratrix L. Assessing topic model relevance: Evaluation and informative priors // *Statistical Analysis and Data Mining: The ASA Data Science Journal*. — 2019. — Vol. 12, No. 3. — P. 210–222.

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. — P. 132–144.

Hoffman M. D., Blei D. M., Bach F. R. Online learning for latent Dirichlet allocation // Neural Information Processing Systems (NIPS). — Curran Associates, Inc., 2010. — P. 856–864.

Hofmann T. Probabilistic latent semantic indexing // Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. — NY: ACM, 1999. — P. 50–57.

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. / eds. by S. Balandin, V. Niemi, T. Tutina. — 2019. — P. 131–138.

Kochedykov D. A., Apishev M. A., Golitsyn L. V., Vorontsov K. V. Fast and modular regularized topic modelling // Proceeding of the 21st Conference of Finnish-Russian University Cooperation in Telecommunications Association. The seminar on Intelligence, Social Media and Web. Helsinki, Finland, November 6–10, 2017. — IEEE, 2017. — P. 182–193.

Lau J. H., Newman D., Baldwin T. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality // Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics. — 2014. — P. 530–539.

Litvak M., Vanetik N., Liu C., Xiao L., Savas O. Improving summarization quality with topic modeling // Proceedings of the 2015 Workshop on Topic Models: Post-Processing and Applications. — NY: Association for Computing Machinery, 2015. — P. 39–47.

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: Association for Computational Linguistics, 2010. — P. 100–108.

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. — NY: ACM, 2010. — P. 215–224.

Riedl M., Biemann C. TopicTiling: A text segmentation algorithm based on LDA // Proceedings of ACL 2012 Student Research Workshop. — ACL '12. — Stroudsburg, PA, USA: Association for Computational Linguistics, 2012. — P. 37–42.

Röder M., Both A., Hinneburg A. Exploring the space of topic coherence measures // Proceedings of the eighth ACM international conference on Web search and data mining / ACM. — 2015. — P. 399–408.

Rubin T. N., Chambers A., Smyth P., Steyvers M. Statistical topic models for multi-label document classification // *Machine Learning*. — 2012. — Vol. 88, No. 1-2. — P. 157–208.

Skachkov N. A., Vorontsov K. V. Improving topic models with segmental structure of texts // Proceedings of 24-th Conference «Dialogue» on Computational Linguistics and Intellectual Technologies. — 2018. — P. 652–661.

Sokolov E., Bogolubsky L. Topic models regularization and initialization for regression problems // Proceedings of the 2015 Workshop on Topic Models: Post-Processing and Applications. — NY: ACM, 2015. — P. 21–27.

Taddy M. On estimation and selection for topic models // Artificial Intelligence and Statistics. — 2012. — P. 1184–1193.

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 // Analysis of Images, Social networks and Texts. — Springer International Publishing Switzerland, Communications in Computer and Information Science (CCIS), 2015. — P. 370–384.

Vorontsov K. V., Potapenko A. A. Tutorial on probabilistic topic modeling: Additive regularization for stochastic matrix factorization // Analysis of Images, Social networks and Texts. — Vol. 436. — Springer International Publishing Switzerland, Communications in Computer and Information Science (CCIS), 2014. — P. 29–46.

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. — P. 303–323.

Yanina 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 / eds. by A. Filchenkov, L. Pivovarova, J. Žižka. — Springer International Publishing, Cham, 2018. — P. 181–193.