Научная статья
УДК 004.94:008.2
https://doi.org/10.35266/1999-7604-2024-3-7

Оптимизация при вероятностном тематическом моделировании технологической прогностической информации

Олег Русланович Попов[1]✉, Сергей Олегович Крамаров[2]
[1]Академия информатизации образования, Ростов-на-Дону, Россия
[2]Сургутский государственный университет, Сургут, Россия
[1]cs41825@aaanet.ru✉, https://orcid.org/0000-0001-6209-3554
[2]kramarov_so@surgu.ru, https://orcid.org/0000-0003-3743-6513

Аннотация. На основе анализа методов мягкой кластеризации документов и вероятностных распределений терминов и тем рассмотрены вычислительные методы и инструменты моделирования динамики политематических потоков в многомерном информационном пространстве. Предложена оптимизированная стохастическая модель динамики мягкой кластеризации сетей знаний в информационном пространстве, структурированном на основе семантических связей в текстах определенной предметной области, извлеченных из наукометрических и библиографических баз данных. На основе теоретической модели разработан алгоритм и методика его применения, с помощью которой возможно применение расширенной текстовой аналитики, включая выявление скрытых тем и прогнозирование трендов. Разработанная методика позволяет с определенным уровнем научной объективности осуществлять прогнозирование новых технологий и актуальных научных направлений в заданной определенной предметной исследовательской области, в том числе для решения теоретических, прикладных и управленческих задач. На основании практических результатов, полученных в работе, разработан глоссарий прогностических терминов «Информационные технологии и коммуникации», который рекомендован к применению в учебном процессе системы общего и профессионального образования.

Ключевые слова: тематическая модель, вероятность, алгоритм, стохастическая модель, информация, семантика, текст, прогностический термин, информационно-коммуникационные технологии

Для цитирования: Попов О. Р., Крамаров С. О. Оптимизация при вероятностном тематическом моделировании технологической прогностической информации // Вестник кибернетики. 2024. Т. 23, № 3. С. 56–69. https://doi.org/10.35266/1999-7604-2024-3-7.

Original article

Optimization in probabilistic topic modeling of technological predictive information

Oleg R. Popov[1]✉, Sergey O. Kramarov[2]
[1]Academy of Informatization of Education, Rostov-on-Don, Russia
[2]Surgut State University, Surgut, Russia
[1]cs41825@aaanet.ru✉, https://orcid.org/0000-0001-6209-3554
[2]kramarov_so@surgu.ru, https://orcid.org/0000-0003-3743-6513

Abstract. The analysis of soft clustering methods of documents and probabilistic distributions of terms and topics leads us to consider computational methods and tools for modeling the dynamics of polytopic flows in a multidimensional information space. We propose an optimized stochastic model that captures the dynamics of soft clustering of knowledge networks in an information space. This model is based on semantic connections in texts of a specific subject area, which are extracted from scientometric and bibliographic databases. Using the theoretical model, we developed an algorithm and methodology for applying advanced text analytics, which includes the identification of hidden topics and the prediction of trends. The developed methodology

allows, with a certain level of scientific objectivity, to predict new technologies and current scientific directions in a given specific research area, including for solving theoretical, applied and management problems. The practical results led to the development of a glossary of predictive terms "Information Technologies and Communications". We recommend using this glossary in the educational process of the general and vocational education system.

Keywords: topic model, probability, algorithm, stochastic model, information, semantics, text, predictive term, information and communication technologies

For citation: Popov O. R., Kramarov S. O. Optimization in probabilistic topic modeling of technological predictive information. *Proceedings of Cybernetics*. 2024;23(3):56–69. https://doi.org/10.35266/1999-7604-2024-3-7.

ВВЕДЕНИЕ

Развитие стохастического анализа доказало важность подхода, основанного на стохастическом описании природы многих сложных физических и технологических систем. Наблюдается расширение использования вероятностных методов и для прогнозирования развития научных областей. В настоящее время сформировался целый блок научно-технических дисциплин, имеющих общую системную ориентацию, задающую относительно них особую плоскость существования искусственно создаваемых сложных систем. Особый интерес представляют междисциплинарные области взаимных пересечений наук и технологий. На этих стыках используются инструменты одной области для продвижения другой.

Наиболее развитыми в динамике процесса представляются информационно-коммуникационные технологии (ИКТ). Чаще всего именно эти технологии поставляют инструменты для развития других технологий через возможности имитационного компьютерного моделирования различных процессов.

Для описания процессов, учитывающих сложные нелинейные механизмы, необходимо использование специального подхода, связанного с вероятностным тематическим моделированием. Вероятностные тематические модели осуществляют «мягкую кластеризацию» (soft clustering), относя документ к нескольким кластерам-темам с некоторыми вероятностями.

Проблема вероятностного моделирования динамики мягкой кластеризации тематических информационных потоков, особенно с учетом внутри- и межструктурных взаимодействий, а также многих метаданных (модальностей), признаются открытыми научными проблемами [1–3].

Актуальной практической задачей является проблема моделирования стохастических процессов взаимодействия политематических информационных потоков, структурированных на основе семантических связей определенной предметной области, для прогнозирования новых технологий и научных направлений. В качестве предметной области выбраны прогнозируемые ИКТ, которые находятся в кластерах пересечения научных инноваций и дают рекомендации для исследований по другим зрелым или появляющимся технологиям [4].

МАТЕРИАЛЫ И МЕТОДЫ

Основные вычислительные методы и инструменты, приводящие к автоматическому режиму обнаружения знаний и скрытых ассоциаций в публикациях, включая тематическое моделирование, подробно классифицируются в источнике [5]. Анализ на уровне темы распространения дополняет процесс поиска и фильтрации информации, помимо анализа на уровне термов.

Тематические модели используются в различных предметных областях и для различных задач, включая, например, менеджмент знаний [6], анализ кластеризации сообществ [7], автоматическое выявление новых новостных тем [8]. Достаточно редки публикации, отражающие использование данных методов для анализа научно-технологических инноваций.

Аналитический метод, позволяющий сгруппировать связанные термины и фразы, определить меняющиеся тематические акценты

в сфере информатики и больших данных, составить технологическую дорожную карту (technology roadmap, TRM) предложен в источнике [9]. TRM прогнозирует будущие изменения в технологических темах и получает информацию для планирования научно-исследовательских и опытно-конструкторских работ, а также для стратегического управления. Однако данная работа представляет собой гибридный набор различных методик при отсутствии явной вычислительной модели. Для прогнозирования применяется сложный в реализации экспертный подход без автоматизации данных. Двумерное отображение семантической близости тем в виде «дорожной карты» является весьма упрощенным инструментом визуализации [10].

В источнике [11] на основе анализа обширной коллекция статей, опубликованных с 2000 по 2021 гг. на конференциях по машинному обучению и искусственному интеллекту (ИИ), реализуется задача ранней детекции трендовых научных тем. Обосновывается преимущество предлагаемого инкрементального метода вероятностного тематического моделирования, реализуемого на основе модели ARTM, в сравнении с популярными байесовскими и нейросетевыми подходами. В качестве примеров выявленных трендовых тем в машинном обучении приводятся «LSTM», «deep learning», «word2vec», «BERT», «fake news detection». Однако очевидные прикладные результаты работы модели не приводятся.

Потенциальной возможностью вероятностных подходов для решения актуальных прикладных научных задач в сфере технологий ИИ является информация о том, что в лаборатории машинного обучения и семантического анализа Института искусственного интеллекта МГУ обучили нейронную сеть для получения семантических векторных представлений (эмбеддингов) научных текстов на русском языке SciRus-tiny [12]. Это реализовано с помощью данных, предоставленных для обучения порталом eLIBRARY.RU, который содержит большое количество документов по множеству разнообразных научных тематик.

Полученные результаты показывают, что сжатое описание документа в виде вектора вероятностей тем содержит важнейшую информацию о семантике документа и может использоваться для решения многих нетривиальных задач текстовой аналитики, включая обнаружение латентной кластерной структуры в документах и выявление трендов в библиографических и патентных базах данных.

Вероятностная тематическая модель (probabilistic topic model) выявляет тематику коллекции документов, представляя каждую тему дискретным распределением вероятностей терминов, а каждый документ – дискретным распределением вероятностей тем.

Исходными данными для тематического моделирования является множество (коллекция) текстовых документов D и множество (словарь) терминов (термов) W. Под термами понимаются слова, нормальные формы слов, словосочетания или термины, в зависимости от того, какие виды предварительной обработки текстов были выполнены.

Каждый документ $d \in D$ представляется последовательностью термов $(w_1, ..., w_{n_d})$ из W, где n_d – длина документа. Через n_{dw} обозначается число вхождений терма w в документ d.

Существует конечное множество тем T и коллекция порождается дискретным распределением $p(d,w,t)$ на $D \times W \times T$. Документы d и термы w являются наблюдаемыми переменными, тема t – латентной переменной, т. е. каждое вхождение терма w в документ d связано с некоторой неизвестной темой t из заданного конечного множества T.

Для каждой темы t и документа d зададим вероятность темы в документе $p(t|d)$. То же самое сделаем для слов и тем: $p(w|t)$ – вероятность встретить слово w в теме t. Распределение вероятностей термов $p(w|d,t)$ зависит только от темы t, но не от документа d (гипотеза условной независимости):

$$p(w|d,t) = p(w|t). \quad (1)$$

Распределение терминов в документе $p(w|d)$ описывается вероятностной смесью распределений терминов в темах $\varphi_{\omega t} = p(w|t)$ с весами $\theta_{td} = p(t|d)$:

$$p(w|d) = \sum_{t\in T} p(w|d,t)p(t|d) = \\ = \sum_{t\in T} p(w|t)p(t|d) = \sum_{t\in T} \varphi_{\omega t}\, \theta_{td.} \qquad (2)$$

Вероятностная порождающая или генеративная модель (2) описывает процесс порождения коллекции по известным распределениям $p(w|t)$ и $p(t|d)$.

Построить тематическую модель коллекции является обратной задачей [13], что означает найти по заданной коллекции D множество тем T, условные распределения термов $\varphi_{\omega t} = p(w|t)$ для каждой темы $t \in T$ с весами $\theta_{td} = p(t|d)$ для каждого документа $d \in D$.

Равенство (2) можно представить в виде матрицы $P = (p_{dw})_{W\times D}$ частот слов документа (как часто каждое слово встречается в каждом документе), которая затем записывается как произведение двух матриц меньших размеров $\Phi = (\varphi_{\omega t})_{W\times T}$ – матрицы термов тем и $\Theta = (\theta_{td})_{T\times D}$ – матрицы тем документов. Матрица P известна, так как это исходные данные, и имеет в общем случае полный ранг, поэтому не может быть в точности равна $\Phi\Theta$. Правая часть равенства представляет собой произведение двух неизвестных матриц. Построение вероятностной тематической модели является решением задачи поиска приближенного низкорангового стохастического матричного разложения $P \approx \Phi\Theta$ (рис. 1).

Столбцы этих матриц в целом можно рассматривать как нормализованные, то есть задача декомпозиции матрицы может быть приведена к максимизации логарифма правдоподобия [14]:

$$\sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{t\in T} \varphi_{wt}\theta_{td} \;\to\; \max_{\Phi,\Theta} \qquad (3)$$

при ограничениях неотрицательности и нормировки всех столбцов φ_t, θ_d:

$$\sum_{w\in W} \varphi_{wt} = 1; \quad \varphi_{wt} \geqslant 0;$$

$$\sum_{t\in T} \theta_{td} = 1; \quad \theta_{td} \geqslant 0. \qquad (4)$$

Решение задачи матричного разложения будет неуникальным. Согласно теории регуляризации А. Н. Тихонова [15], решение некорректно поставленной операторной задачи возможно доопределить и сделать устойчивым. Для этого следует наложить некоторые дополнительные ограничения на $R(\Phi, \Theta)$, называемые регуляризаторами. Функция регуляризатора должна быть непрерывно дифференцируемой. К задаче максимизации логарифмического правдоподобия будет до-

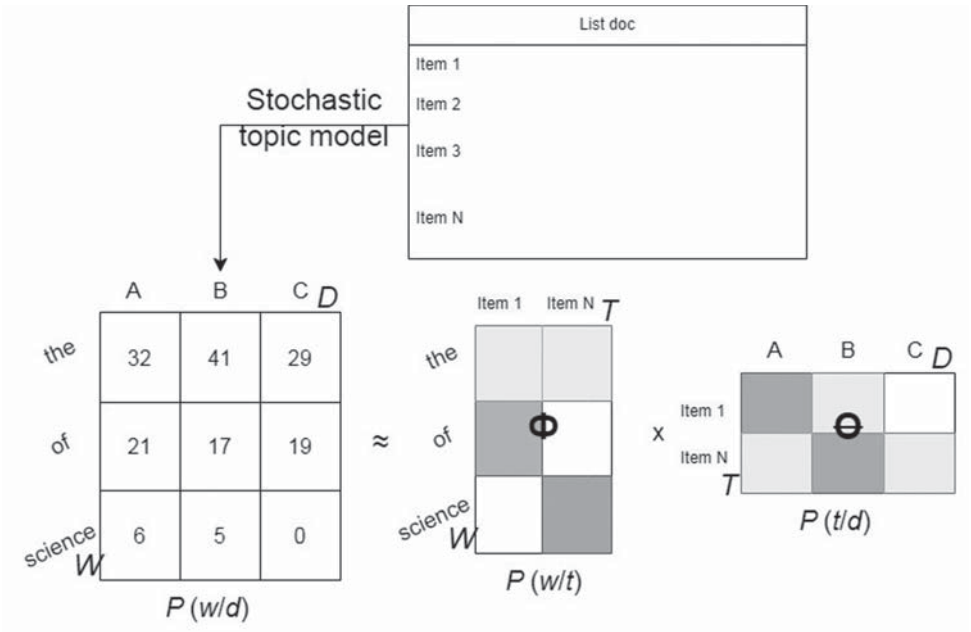

Рис. 1. Построение тематической модели: решение задачи приближенного стохастического матричного разложения $P \approx \Phi\Theta$

Примечание: составлено авторами на основании данных, полученных в исследовании.

бавлен новый компонент, и задача оптимизации примет вид:

$$\sum_{d\in D}\sum_{\omega\in d} n_{d\omega}\ln\sum_{t\in T}\phi_{\omega t}\theta_{td} + \\ + R(\Phi,\Theta)\to\max_{\Phi,\Theta}, \tag{5}$$

Первая модель, называемая pLSA (вероятностный латентный семантический анализ), предложенная в источнике [16], предполагала, что $R(\Phi, \Theta) = 0$.

В дальнейшем широкое применение получили два подхода к регуляризации вероятностных генеративных моделей: скрытое распределение Дирихле (latent Dirichlet allocation, LDA) [17], основанный на байесовском выводе, и, более классическая, аддитивная регуляризация тематических моделей, которая, как следует из названия, лежит в основе ARTM.

Модель LDA предполагает байесовскую регуляризацию таким образом, что:

$$\sum_{d\in D}\sum_{t\in T}(\alpha_t-1)\ln\theta_{td}. \\ +\sum_{d\in D}\sum_{t\in T}(\alpha_t-1)\ln\theta_{td}. \tag{6}$$

Здесь β_ω и α_t являются гиперпараметрами с положительной настройкой. Модель LDA предполагает, что векторы документа θ_{td} генерируются одним и тем же распределением вероятностей на нормализованных векторах, а распределение берется из семейства распределений Дирихле с параметром α [7]. Аналогично, векторы темы $\varphi_{\omega t}$ генерируются распределением Дирихле с параметром β.

Распределение Дирихле существенно упрощает байесовский вывод, и большинство моделей строятся с его использованием.

Однако следует обратить внимание на ряд проблем, связанных с байесовской регуляризацией: необходимо оптимизировать гиперпараметры, инициализировать параметр β_ω для новых терминов и обеспечивать разреженность, при том, что обнулять параметры $\varphi_{\omega t}$ и θ_{td} невозможно.

Введение других инструментов приближенного байесовского вывода (вариационный вывод, семплирование Гиббса, распространение ожидания) не позволяет легко комбинировать модели и снимать ограничения, связанные с выбором распределений Дирихле. Для каждой новой модели приходится заново выполнять математические выкладки и программную реализацию [13].

Альтернативой байесовскому подходу является метод аддитивной регуляризации тематических моделей (ARTM) [18]. Это приложение классической теории регуляризации некорректно поставленных задач [15] к тематическому моделированию.

Аддитивная регуляризация тематических моделей (ARTM) основана на максимизации линейной комбинации логарифма правдоподобия и нескольких регуляризаторов $R_i(\Phi, \Theta)$, $i = 1, \ldots, k$:

$$\sum_{d\in D}\sum_{w\in d} n_{dw}\ln\sum_{t\in T}\varphi_{wt}\theta_{td} + \\ +\underbrace{\sum_{i=1}^{k}\tau_i R_i(\Phi,\Theta)}_{R(\Phi,\Theta)}\to\max_{\Phi,\Theta}; \tag{7}$$

при прежних ограничениях (4), где τ_i – неотрицательные коэффициенты регуляризации.

Построение многофункциональных тематических моделей существенно упрощается благодаря аддитивности регуляризаторов [18]. Функционал максимизации правдоподобия (3) позволяет добавить не один метод регуляризации, а несколько, придавая вес каждому. Этот подход выражает суть аддитивности регуляризации и, как следует из названия, лежит в основе ARTM.

Таким образом, ARTM это не инкрементное улучшение одной тематической модели, а общий подход к тематическому моделированию как к задаче многокритериальной оптимизации.

В подходе APTM распределение Дирихле является не универсальным, а одним из возможных регуляризаторов. В качестве базовой модели логичнее брать pLSA, не имеющую собственных регуляризаторов, которые можно добавлять из модульной расширяемой библиотеки в зависимости от поставленной проблемы.

С точки зрения поставленных в работе прикладных исследовательских задач, такой подход дает преимущества. В первую очередь это

касается оптимизации стратегий построения иерархических тематических моделей, включая динамические модели, которые выявляют закономерности развития кластеров не только внутри мультидисциплинарных корпусов, но и с течением времени.

При формировании базовых принципов построения многофункциональной тематической модели (далее «Модель») необходимо учитывать максимум дополнительной информации, особенности семантики и предмета текстовой коллекции. С точки зрения стратегий оптимизации при вероятностном тематическом моделировании выделяются следующие факторы:

1. Автоматическая генерация графов.

2. Ранжирование при оценке результатов поиска.

3. Учет отсутствующей в классических алгоритмах ранжирования силы связей между словами и предложениями и других семантических метрик.

4. Учет при моделировании сетевой топологии показателя центральности, содержащего необходимый и достаточный объем информации, и показателя промежуточности, определяющего степень передачи информации без потерь и искажений.

5. Привязка документа ко времени создания для оценки динамики развития событий, выявления закономерностей и прогнозирования.

6. Учет при построении модели точек сближения автоматической суммаризации и тематического подхода к обработке коллекций документов.

Общая схема многофункциональной модели обработки информационных процессов, рассматриваемой как большой научный проект science data, представлена на рис. 2. В Модели выделяются три блока, характерные также для разработки и проектирования проекта big data с использованием технологий глубинного анализа текстов (Text Mining) и нахождения при этом закономерностей и трендов:

1. Подготовка и упорядочение данных.

2. Импорт данных, их обработка и генерация новой информации.

3. Валидация и оценка качества результатов моделирования.

В рамках данных блоков сформированы пять взаимосвязанных этапов алгоритма Модели:

1. Предварительная информационная экспертиза.

2. Формирование понятийного ядра предметной области и концептуальных тематических запросов.

3. Автономный сбор и накопление метаданных коллекций документов.

4. Многофункциональная обработка данных, генерация и визуализация информации.

5. Контроль, оценка и валидация данных моделирования.

Заданный алгоритмический подход преследует цель обеспечить максимальную объ-

Рис. 2. Общая схема моделирования обработки информационных процессов как научного проекта science data

Примечание: составлено авторами на основании данных, полученных в исследовании.

ективность поиска и скорость, с которой он позволяет углубиться в выбранную предметную область исследований.

РЕЗУЛЬТАТЫ И ИХ ОБСУЖДЕНИЕ

На первом подготовительном этапе практической реализации Модели используется авторская методика расчета показателя уровня зрелости самоорганизующихся интеллектуальных систем «SML», интегрирующего в себе показатели «технологические уровни готовности» (TPL) и «уровни социоэкономической адаптированности» (SPL) прогнозируемых технологий, входящих в исследуемую социотехническую систему [19].

В целях определения базовых структур предметной области экспертным методом выделены категории перспективных направлений развития ИКТ, соответствующих определенному уровню SML-матрицы. Например, в таблице приведены выделенные в рамках 4 категорий 22 перспективных направления развития информационно-коммуникационных технологий, соответствующих четвертому уровню:

1. Человеко-машинные интерфейсы (human-computer interfaces);
2. Инженерия вычислений (computing engineering);
3. Технологии памяти и хранения данных (memory and data storage technologies);
4. Электроника и коммуникации (electronics and communications) [20].

Далее, исходя из предложенного подхода к формированию понятийного ядра предметной области, разработан алгоритм, применимый к задаче формирования «словаря прогностических терминов» [21].

Данный алгоритм позволяет на базе внешних сервисов (например, «Википедии») формировать словарь контекстно-близких терминов по отношению к изначально заданному термину. При этом задается набор изначальных терминов, и по результатам их обработки

Таблица

Категории перспективных направлений развития ИКТ, соответствующих четвертому уровню SML-матрицы

I	**Human – computer interfaces**	
1	ambient intelligence	интеллектуальная среда
2	brain – computer interface	интерфейс «мозг– компьютер»
3	city brain	городской мозг
4	semantic web	семантический веб
5	smart city	умный город
II	**Computing engineering**	
1	exascale computing	масштабные вычисления
2	neuromorphic engineering	нейроморфная инженерия
3	optical computing	оптические (фотонные) вычисления
4	quantum computing	квантовые вычисления
III	**Memory and data storage technologies**	
1	3D optical data storage	3D оптическое хранение данных
2	DNA digital data storage	цифровое хранилище данных ДНК
3	holographic data storage	голографическое хранилище данных
4	patterned media	узорчатые носители
5	phase-change memory	память с фазовым переходом
6	quantum memory	квантовая память
IV	**Electronics and communications**	
1	atomtronics	атомтроника
2	carbon nanotube field-effect transistor	полевой транзистор из углеродных нанотрубок
3	Li-Fi (Light Fidelity)	Li-Fi
4	memristor	мемристор, мемтранзистор, мемистор, транситор
6	software-defined radio	программно-определяемое радио
7	spintronics	спинтроника, твистроника, валлейтроника

Примечание: составлено по источнику [21].

формируется полноценный словарь. Алгоритм оценивает «вес» близости терминов по отношению к изначальному термину. Ранжирование терминов выполняется по средней арифметической сумме оценок алгоритмов PageRank и HITS.

В процессе работы алгоритма набор собранных терминов группируется по категориям технологий, заданных программой исследования.

На рис. 3 представлена демонстрация работы алгоритма на примере графа, созданного по отношению к первичному термину «quantum computing» (квантовые вычисления).

По результатам работы алгоритма могут быть выведены различные варианты сводных списков терминов, как общих, так и структурированных по категориям технологий.

На основе глоссария прогностических терминов сформирована база концептуальных тематических запросов в сети знаний, представляющих актуальные общедоступные наукометрические и библиографические базы данных (ББД), научные и новостные научно-технические порталы, тематические сетевые ресурсы, тезаурусы.

Уровень достоверности получаемой информации может быть повышен за счет формирования комбинированного алгоритма сбора коллекции текстовых данных из высокоавторитетных качественных исследовательских баз данных (БД), таких как Google Scholar, arXiv, Springer Link, eLIBRARY.Ru, CyberLeninka, за определенный период времени.

Наукометрические индикаторы, обработанные в выбранной библиографической БД по специальной формуле, представляют собой информационную основу для автономного сбора и накопления метаданных коллекции.

Для извлеченных публикаций автоматически выгружаются их полные библиографические описания в форматах .csv и .xlsx для дальнейшей обработки и интеллектуального текстового анализа. По каждой записи может быть выдана информация, включающая полные «выходные данные», количество цитируемых в публикации источников, различные вторичные поля, несущие справочную

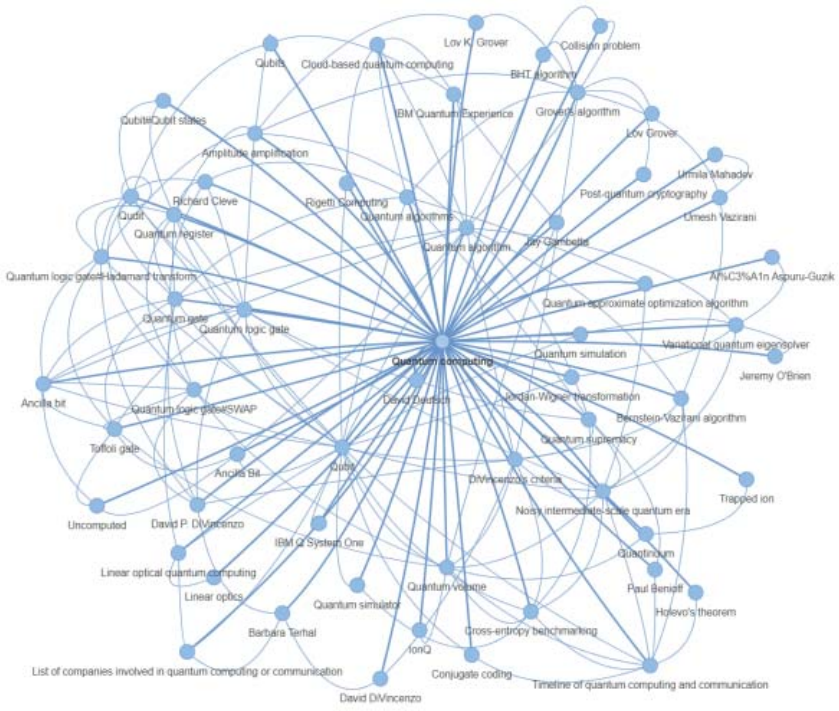

Рис. 3. Визуализация графа терминов/словосочетаний, полученных на основе входного термина «quantum computing»

Примечание: составлено по источнику [21].

информацию. Фрагмент основных информационных полей (Global Number, Time, Title, Authors, Abstract) библиографических данных по запросу «quantum computing» в ББД Google Scholar показан на рис. 4.

Сбор документов осуществляется в автоматическом режиме, с обработкой полученных исходных данных в соответствии с определенными правилами и сохранением метаданных. Выходные данные препроцессора документа представляют собой набор «чистого» текстового контента со связанными метаданными.

Обработанные текстовые документы передаются на вход многофункциональной системы обработки данных, генерации и визуализации информации.

Модуль тематического моделирования (BigARTM, Gensim-LDA) автоматически выводит набор скрытых тем. Результаты анализа темы, включая ключевое слово темы, дистрибутивы и распределения по темам документов, также сохраняются и индексируются во время индексации тем.

В результате работы алгоритма на данном этапе темпоральная (датированная) коллекция текстовых документов систематизируется по двум уровням:

- в соответствии с базовыми концепциями тематического запроса;

- в соответствии с проведенным тематическим анализом документов, сформированных внешней ББД в ответ на запрос.

Автономный и сетевой режимы ориентированы на решение как фундаментальных исследовательских задач, так и прикладных интересов пользователей информационного ресурса в процессе активного взаимодействия с системой. В данном основном блоке реализованы базовые процессы обработки big data, генерации новой информации и ее визуального представления с целью нахождения при этом закономерностей и прогностических трендов.

Чтобы отслеживать появление новых тем, для каждого временного шага в ББД обновляются данные. При поступлении новой порции документов D' словарь пополняется новыми терминами W' и могут образоваться новые темы T'. Основной принцип выявления научных трендов заключается в том, что новая лексика, появившаяся в новых документах, относится преимущественно к новым темам [11].

Темпоральное исследование кластерообразования, связанного с научной областью квантовых вычислений (quantum computing) в период 1980–2023 гг., приведено на рис. 5.

Анализ динамики тематических данных, интегрированных суммарно в 10 кластеров, показывает активное кластерообразование в новейший период (2020–2023 гг.) в области кластеров C6 («новые технологии»), C7 («квантовая нейронная сеть») и C10 («квантовое шифрование изображения») (рис. 3). Так, область кластера C7 пополняется такими терминами, как «квантовая информация», «квантовые данные», «классификация изображений», «процессор», «глубокая нейронная сеть», «сверточная нейронная сеть». В области кластера C10 можно обнаружить «шифрование цветного изображения», «логистическая карта», «схема квантового шифрования изображения», «квантовая репрезентативная модель». Использование новой, специфичной для данной научной области, лексики релевантно отражает постепенное вхождение новых квантовых информационных технологий, новых архитектур квантовых процессоров и симуляторов, устройств передачи квантовой информации, отражающих сложную динамику новейших квантовых систем и коммуникаций.

Особенностью выбранной методики валидации данных является рассмотрение, помимо классических мер оценки тематического моделирования (перплексии, когерентности, разреженности), критериев, одновременно выступающих индикаторами развития исследовательской модели.

Исходя из выбранного подхода к построению Модели, предложена модифицированная методика количественной оценки релевантности динамически выявленных скрытых тем начальному запросу (ключевому слову или концепту) [22].

Критерием релевантности темы понятию является существование между ними взаимно однозначного соответствия. Для выявле-

Global_Number	Time	Title	Authors	Abstract
513	07.01.2022	Toward implementing efficient image processing algorithms on quantum computers	['Fei Yan', 'Salvador E. Venegas-Andraca', 'Kaoru Hirota']	Quantum information science is an interdisciplinary subject spannin
527	10.02.2022	Quantum computational complexity from quantum information to black holes and back	['Shira Chapman', 'Giuseppe Policastro']	Quantum computational complexity estimates the difficulty of cons
528	06.08.2020	Hybridization of Moth flame optimization algorithm and quantum computing for gene selection in m	['Ali Dabba', 'Abdelkamel Tari', 'Samy Meftali']	Ever-increasing data in various fields like Bioinformatics field, which
529	02.09.2021	Quantum federated learning through blind quantum computing	['Weikang Li', 'Sirui Lu', 'Dong-Ling Deng']	Private distributed learning studies the problem of how multiple dist
530	04.09.2021	Balanced k-means clustering on an adiabatic quantum computer	['Davis Arthur', 'Prasanna Date']	Adiabatic quantum computers are a promising platform for efficient
531	31.08.2021	Unsupervised event classification with graphs on classical and photonic quantum computers	['Andrew Blance', 'Michael Spannowsky']	Photonic Quantum Computers provide several benefits over the dis
533	16.03.2021	Quantum Computing and Deep Learning Methods for GDP Growth Forecasting	['David Aiaminos', 'M. Belén Salas', 'Manuel A. Fernández-Gámez']	Precise macroeconomic forecasting is one of the major aims of ecor
534	22.05.2020	Designing nanotechnology QCA–multiplexer using majority function-based NAND for quantum comp	['Jun-Cheol Jeon']	Quantum-dot cellular automata (QCAs) are one of the most signific
535	17.02.2022	Collective neutrino oscillations on a quantum computer	['Kübra Yeter-Aydeniz', 'Shikha Bangar', 'George Siopsis', 'Raphael C. Pooser']	We calculate the energy levels of a system of neutrinos undergoing
536	20.07.2021	Toward simulating superstring/M-theory on a quantum computer	['Hrant Gharibyan', 'Masanori Hanada', 'Masazumi Honda', 'Junyu Liu']	We present a novel framework for simulating matrix models on a q
537	07.03.2022	Constant-depth circuits for dynamic simulations of materials on quantum computers	['Lindsay Bassman Oftelie', 'Roel Van Beeumen', 'Ed Younis', 'Ethan Smith', 'Costin Iancu', 'Wibe A. c	Dynamic simulation of materials is a promising application for near-
538	25.01.2021	Methods for accelerating geospatial data processing using quantum computers	['Max Henderson', 'Jared Gallina', 'Michael Brett']	Quantum computing is a transformative technology with the poten
542	18.06.2021	Compiling basic linear algebra subroutines for quantum computers	['Liming Zhao', 'Zhikuan Zhao', 'Patrick Rebentrost', 'Joseph Fitzsimons']	Efficiently processing basic linear algebra subroutines is of great im
544	04.08.2021	Hybrid quantum–classical convolutional neural networks	['Junhua Liu', 'Kwan Hui Lim', 'Kristin L. Wood', 'Wei Huang', 'Chu Guo', 'He-Liang Huang']	Deep learning has been shown to be able to recognize data pattern
546	27.09.2022	NISQ computing: where are we and where do we go?	['Jonathan Wei Zhong Lau', 'Kian Hwee Lim', 'Harshank Shrotriya', 'Leong Chuan Kwek']	In this short review article, we aim to provide physicists not working
551	22.01.2022	Unconstrained binary models of the travelling salesman problem variants for quantum optimization	['Özlem Salehi', 'Adam Glos', 'Jarosław Adam Miszczak']	Quantum computing is offering a novel perspective for solving com
552	14.05.2021	Review of Quantum Image Processing	['Zhaobin Wang', 'Minzhe Xu', 'Yaonan Zhang']	As an interdisciplinary between quantum computing and image pro
553	10.01.2022	Emerging Enabling Technologies for Industry 4.0 and Beyond	['Alexander Sigov', 'Leonid Ratkin', 'Leonid A. Ivanov', 'Li Da Xu']	Rapid advances in technology have spurred tremendous progress ir
554	03.01.2021	Event Classification with Quantum Machine Learning in High-Energy Physics	['Koji Terashi', 'Michiru Kaneda', 'Tomoe Kishimoto', 'Masahiko Saito', 'Ryu Sawada', 'Junichi Tanaka	We present studies of quantum algorithms exploiting machine learr
555	08.03.2021	Cloud based QC with Amazon Braket	['Constantin Gonzalez']	The approaches to solutions in quantum computing technology are
556	31.05.2022	Quantum k-means clustering method for detecting heart disease using quantum circuit approach	['S S Kavitha', 'Narasimha kaulgud']	The development of noisy intermediate- scale quantum computers
557	24.02.2021	Quantum machine learning for particle physics using a variational quantum classifier	['Andrew Blance', 'Michael Spannowsky']	Quantum machine learning aims to release the prowess of quantun
560	17.03.2021	Quantum beetle antennae search: a novel technique for the constrained portfolio optimization probl	['Ameer Tamoor Khan', 'Xinwei Cao', 'Shuai Li', 'Bin Hu', 'Vasilios N. Katsikis']	In this paper, we have formulated quantum beetle antennae searcl
561	08.06.2021	Highlighting photonics: looking into the next decade	['Zhigang Chen', 'Mordechai Segev']	Let there be light–to change the world we want to be! Over the pas
562	18.07.2023	Design of Quantum Communication Protocols in Quantum Cryptography	['Bilal A. Alhayani', 'Omar A. Alkawak', 'Hemant B. Mahajan', 'Haci Ilhan', 'Roa'a Mohammed Qase	Secure communication has developed into one of the most promisi
563	07.09.2022	Quantum Error Correction: Noise-Adapted Techniques and Applications	['Akshaya Jayashankar', 'Prabha Mandayam']	The quantum computing devices of today have tens to hundreds of
565	30.11.2021	Quantum Support Vector Machines for Continuum Suppression in B Meson Decays	['Jamie Heredge', 'Charles Hill', 'Lloyd Hollenberg', 'Martin Sevior']	Quantum computers have the potential to speed up certain comput
566	13.04.2021	Optimizing quantum heuristics with meta-learning	['Max Wilson', 'Rachel Stromswold', 'Filip Wudarski', 'Stuart Hadfield', 'Norm M. Tubman', 'Eleanor G	Variational quantum algorithms, a class of quantum heuristics, are
567	03.07.2021	A Review of Machine Learning Classification Using Quantum Annealing for Real-World Applications	['Rajdeep Kumar Nath', 'Himanshu Thapliyal', 'Travis S. Humble']	Optimizing the training of a machine learning pipeline helps in redu
569	10.01.2022	A quantum convolutional neural network on NISQ devices	['Shi-Ju Ran', 'Yan-Hu Chen', 'Zeng-Rong Zhou', 'Gui-Lu Long']	Quantum machine learning is one of the most promising applicatior
570	10.08.2021	Quantum Machine Learning Architecture for COVID-19 Classification Based on Synthetic Data Gener	['Javaria Amin', 'Muhammad Sharif', 'Nadia Gul', 'Seifedine Kadry', 'Chinmay Chakraborty']	COVID-19 is a novel virus that affects the upper respiratory tract, a
571	28.04.2020	Binary quantum-inspired gravitational search algorithm-based multi-criteria scheduling for multi-proc	['Abhijeet Singh Thakur', 'Tarun Biswas', 'Pratyay Kuila']	A quantum-inspired hybrid scheduling technique is proposed for mu
572	08.06.2021	Probabilistic nonunitary gate in imaginary time evolution	['Tong Liu', 'Jin-Guo Liu', 'Heng Fan']	Simulation of quantum materials is a significant application of quan
573	13.03.2021	Quantum K-Nearest-Neighbor Image Classification Algorithm Based on K-L Transform	['Nan-Run Zhou', 'Xia-Xun Liu', 'Yu-Ling Chen', 'Ni-Suo Du']	Enlightened by quantum computing theory, a quantum K-Nearest-N
574	06.09.2021	Quantum simulation of gauge theory via orbifold lattice	['Alexander J. Buser', 'Hrant Gharibyan', 'Masanori Hanada', 'Masazumi Honda', 'Junyu Liu']	We propose a new framework for simulating U(k) Yang–Mills theor
575	30.03.2022	Measurements of Quantum Hamiltonians with Locally-Biased Classical Shadows	['Charles Hadfield', 'Sergey Bravyi', 'Rudy Raymond', 'Antonio Mezzacapo']	Obtaining precise estimates of quantum observables is a crucial ste
576	03.05.2021	Robust implementation of generative modeling with parametrized quantum circuits	['Vicente Leyton-Ortega', 'Alejandro Perdomo-Ortiz', 'Oscar Perdomo']	Although the performance of hybrid quantum–classical algorithms is
577	30.03.2021	Efficient quantum simulation of open quantum dynamics at various Hamiltonians and spectral densit	['Na-Na Zhang', 'Xiao-Yu Chen', 'Ming-Jie Tao', 'Xiao-Yu Deng', 'Wan-Ting He', 'Xiao-Yu Chen', 'Qing-Jun	Simulation of open quantum dynamics for various Hamiltonians an
578	23.07.2021	Quantum secure multi-party summation protocol based on blind matrix and quantum Fourier transfo	['Xin Yi', 'Cong Cao', 'Ling Fan', 'Ru Zhang']	Multi-party quantum summation is the premise for implementing g
580	25.02.2022	Color image encryption algorithm based on hyperchaotic system and improved quantum revolving g	['Xinyuan Wang', 'Yinjie Su', 'Chao Luo', 'Fuzhong Nian', 'Lin Teng']	Because quantum computing can break encryption systems based

Рис. 4. Фрагмент вывода библиографического описания по запросу «quantum computing»

Примечание: составлено авторами на основании данных, полученных в исследовании.

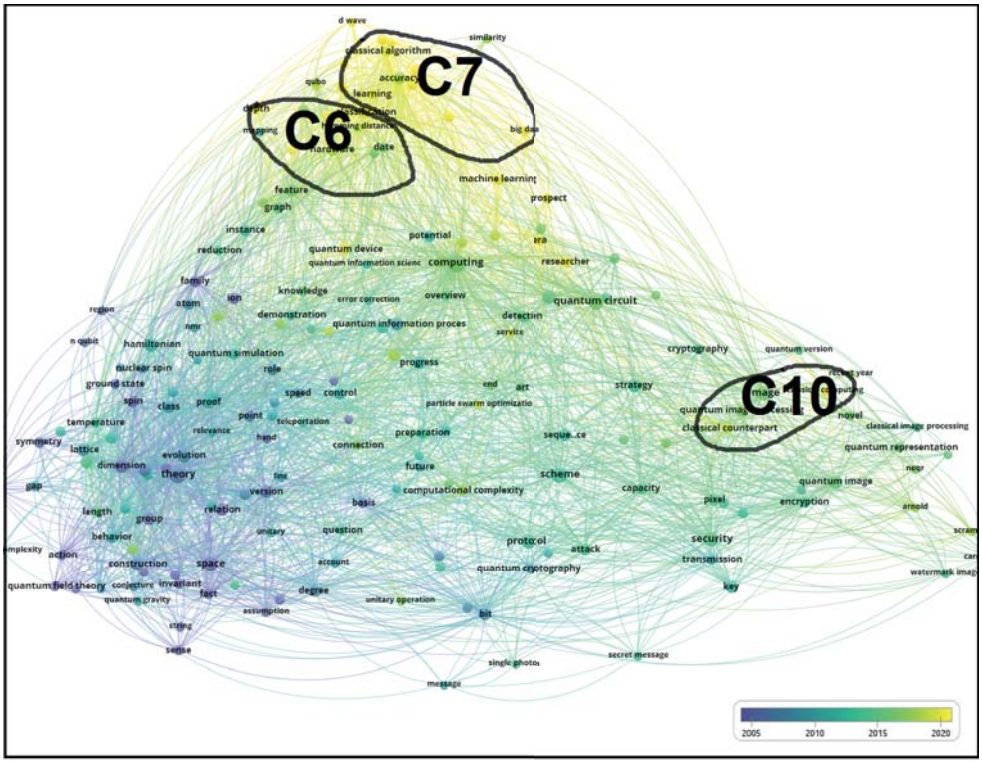

Рис. 5. Карта визуализации VOSviewer, показывающая динамику кластеризации тематических данных в период 1980–2023 гг. в области термина «quantum computing» (квантовые вычисления): цветовая шкала в правом нижнем углу визуализации, отображающая временную оценку данных, варьируется фиолетовым (до 2005 г.), зеленым (период 2010–2015 гг.), желтым (после 2020 г.)

Примечание: составлено авторами на основании данных, полученных в исследовании.

ния степени смещения выявляются 4 типа смещения:

- модель не выявляет темы;
- модель производит смешанные темы;
- ключевые слова в теме отсутствуют;
- повторяются среди скрытых тем.

На первом этапе автоматизированная оценка сходства между темами и концепциями рассчитывается с использованием трех распространенных мер сходства: косинусной, ранговой корреляции Спирмена и KL-дивергенции.

Далее проводится этап сравнительной экспертной оценки совпадений тем и концепций и преобразования оценок сходства в вероятности совпадения.

Составляется матрица размера $n \times m$ всех возможных пар между n ключевыми словами и m выявленными темами. Каждая запись $p(s,t)$ рассматривается как независимая случайная величина Бернулли, представляющая вероятность совпадения того, что эксперт, изучающий распределения соответствия понятия s с темой t, ответит, что они релевантны. Каждой паре $n \times m$ присваивается рейтинг {1, 0,5, 0} для каждого ответа {совпадение, частичное совпадение, отсутствие совпадения}. Пара считается совпадающей, если ее средний рейтинг превышает 0,5.

В результате применения данной методики количественно оценивается и анализируется вероятность соответствия концепта прогнозируемой технологии латентным политематическим потокам, извлеченным и структурированным из научно-технического информационного пространства.

Для дальнейшего анализа качества стохастического моделирования при кластеризации данных и максимизации выявления трендов новых технологий проводится серия экспериментов, где сравнительно рассматриваются вероятностные тематические модели, такие как PLSA, LDA, ARTM [18] с базовыми регуляризаторами матрицы Φ,

и глубокие нейронные сети, в частности BERTopic [11, 12].

В рамках заданной предметной области будут исследованы возможности интеграции тематического моделирования с глубокими нейросетевыми моделями языка, моделями внимания и архитектуры трансформеров.

ЗАКЛЮЧЕНИЕ

1. На основе анализа методов мягкой кластеризации документов и вероятностных распределений терминов и тем рассмотрены вычислительные методы и инструменты моделирования динамики политематических потоков в многомерном информационном пространстве.

2. Предложена оптимизированная стохастическая модель динамики мягкой кластеризации сетей знаний в информационном пространстве, структурированном на основе семантических связей в текстах определенной предметной области, извлеченных из наукометрических и библиографических баз данных.

3. На основе теоретической модели разработан алгоритм и методика его применения, с помощью которой возможно применение расширенной текстовой аналитики, включая выявление скрытых тем и прогнозирование трендов.

4. Разработанная методика позволяет с определенным уровнем научной объективности осуществлять прогнозирование новых технологий и актуальных научных направлений в заданной определенной предметной исследовательской области, в том числе для решения теоретических, прикладных и управленческих задач.

5. На основании практических результатов, полученных в работе, разработан глоссарий прогностических терминов «Информационные технологии и коммуникации», который рекомендован к применению в учебном процессе системы общего и профессионального образования в целях наполнения контента при изучении учебных дисциплин.

6. Проведенное темпоральное исследование кластерообразования, связанное с научной областью квантовых вычислений (quantum computing), отражает сложную динамику новейших квантовых систем и коммуникаций и подтверждает постепенное вхождение новых квантовых информационных технологий.

Список источников

1. Shadrova A. Topic models do not model topics: epistemological remarks and steps towards best practices // Journal of Data Mining & Digital Humanities. 2021. https://doi.org/10.46298/jdmdh.7595.
2. Churchill R., Singh L. The evolution of topic modeling // ACM Computing Surveys. 2022. Vol. 54, no. 10s. P. 1–35. https://doi.org/10.1145/3507900.
3. Zhao H., Phung D., Huynh V. et al. Topic modelling meets deep neural networks: A survey // Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21. 2021. P. 4713–4720. https://doi.org/10.48550/arXiv.2103.00498.
4. Бодрунов С. Д. Ноономика : моногр. М. : Культурная революция, 2018. 432 с.
5. Thilakaratne M., Falkner K., Atapattu T. A systematic review on literature-based discovery: general overview, methodology, & statistical analysis // ACM Computing Surveys. 2019. Vol. 52, no. 6. P. 1–34. https://doi.org/10.1145/3365756.
6. Zelenkov Yu. The topic dynamics in knowledge management research // Knowledge Management in Organizations (KMO 2019): Proceedings of the 14th International Conference. 2019. P. 324–335. https://doi.org/10.1007/978-3-030-21451-7_28.

References

1. Shadrova A. Topic models do not model topics: epistemological remarks and steps towards best practices. *Journal of Data Mining & Digital Humanities*. 2021. https://doi.org/10.46298/jdmdh.7595.
2. Churchill R., Singh L. The evolution of topic modeling. *ACM Computing Surveys*. 2022;54(10s):1–35. https://doi.org/10.1145/3507900.
3. Zhao H., Phung D., Huynh V. et al. Topic modelling meets deep neural networks: A survey. In: *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*. 2021:4713–4720. https://doi.org/10.48550/arXiv.2103.00498.
4. Bodrunov S. D. Noonomika: Monograph. Moscow: Kulturnaya revolyutsiya, 2018. 432 p. (In Russ.).
5. Thilakaratne M., Falkner K., Atapattu T. A systematic review on literature-based discovery: general overview, methodology, & statistical analysis. *ACM Computing Surveys*. 2019;52(6):1–34. https://doi.org/10.1145/3365756.
6. Zelenkov Yu. The topic dynamics in knowledge management research. In: *Proceedings of the 14th International Conference "Knowledge Management in Organizations (KMO 2019)"*. 2019:324–335. https://doi.org/10.1007/978-3-030-21451-7_28.

7. Gorshkov S., Ilyushin E., Chernysheva A. et al. Using topic modeling for communities clusterization in the VKontakte social network // International Journal of Open Information Technologies. 2021. Vol. 9, no. 5. P. 12–17.
8. Zhang J., Ghahramani Z., Yang Y. A probabilistic model for online document clustering with application to novelty detection // Advances in neural information processing systems. 2004. Vol. 17.
9. Zhang Y., Zhang G., Chen H. et al. Topic analysis and forecasting for science, technology and innovation: Methodology with a case study focusing on big data research // Technological forecasting and social change. 2016. Vol. 105. P. 179–191.
10. Айсина Р. М. Обзор средств визуализации тематических моделей коллекций текстовых документов // Машинное обучение анализ данных. 2015. Т. 1, № 11. С. 1584–1618.
11. Герасименко Н. А., Чернявский А. С., Никифорова М. А. и др. Инкрементальное обучение тематических моделей для поиска трендовых тем в научных публикациях // Доклады Российской академии наук. Математика, информатика, процессы управления. 2022. Т. 508, № 1. С. 106–108.
12. Герасименко Н. ruSciBench – бенчмарк для оценки эмбеддингов научных текстов. URL: https://habr.com/ru/articles/781032/ (дата обращения: 25.03.2024).
13. Большакова Е. И., Воронцов К. В., Ефремова Н. Э. и др. Автоматическая обработка текстов на естественном языке и анализ данных. М. : НИУ ВШЭ, 2017. 268 с.
14. Воронцов К. В., Потапенко А. А. Регуляризация, робастность и разреженность вероятностных тематических моделей // Компьютерные исследования и моделирование. 2012. Т. 4, № 4. С. 693–706.
15. Тихонов А. Н., Арсенин В. Я. Методы решения некорректных задач. 3-е изд., испр. М. : Наука, 1986. 286 с.
16. Hofmann T. Probabilistic latent semantic indexing // Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. 1999. P. 50–57.
17. Blei D. M., Ng A. Y., Jordan M. I. Latent dirichlet allocation // Journal of Machine Learning Research. 2003. Vol. 3. P. 993–1022.
18. Воронцов К. В., Потапенко А. А. Аддитивная регуляризация тематических моделей // Доклады Академии наук. 2014. Т. 456, № 3. С. 268–271.
19. Попов О. Р. Адаптация мировых практик к проблеме долгосрочного технологического прогнозирования состояния самоорганизующихся интеллектуальных систем // Интеллектуальные ресурсы – региональному развитию. 2021. № 2. С. 91–98.
20. Крамаров С. О., Попов О. Р., Джариев И. Э. и др. Динамика формирования связей в сетях, структурированных на основе прогностических терминов // Russian Technological Journal. 2023. Т. 11, № 3. С. 17–29. https://doi.org/10.32362/2500-316X-2023-11-3-17-29.

7. Gorshkov S., Ilyushin E., Chernysheva A. et al. Using topic modeling for communities clusterization in the VKontakte social network. *International Journal of Open Information Technologies*. 2021;9(5):12–17.
8. Zhang J., Ghahramani Z., Yang Y. A probabilistic model for online document clustering with application to novelty detection. *Advances in neural information processing systems*. 2004;17.
9. Zhang Y., Zhang G., Chen H. et al. Topic analysis and forecasting for science, technology and innovation: Methodology with a case study focusing on big data research. *Technological forecasting and social change*. 2016;105:179–191.
10. Aysina R. M. Survey of visualization tools for topic models of text corpora. *Machine Learning and Data Analysis*. 2015;1(11):1584–1618. (In Russ.).
11. Gerasimenko N. A., Chernyavskiy A. S., Nikiforova M. A. et al. Inkrementalnoe obuchenie tematicheskikh modeley dlya poiska trendovykh tem v nauchnykh publikatsiyakh. *Doklady Rossijskoj akademii nauk. Matematika, informatika, processy upravleniya*. 2022;508(1):106–108. (In Russ.).
12. Gerasimenko N. ruSciBench – benchmark dlya otsenki embeddingov nauchnykh tekstov. URL: https://habr.com/ru/articles/781032/ (accessed: 25.03.2024). (In Russ.).
13. Bolshakova E. I., Vorontsov K. V., Efremova N. E. et al. Avtomaticheskaya obrabotka tekstov na estestvennom yazyke i analiz dannykh. Moscow: NIU VSHE, 2017. 268 p. (In Russ.).
14. Vorontsov K. V., Potapenko A. A. Regularization, robustness and sparsity of probabilistic topic models. *Computer Research and Modeling*. 2012;4(4):693–706. (In Russ.).
15. Tikhonov A. N., Arsenin V. Ya. Metody resheniya nekorrektnykh zadach. 3d ed., revised. Moscow: Nauka, 1986. 286 p. (In Russ.).
16. Hofmann T. Probabilistic latent semantic indexing. In: *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. 1999:50–57.
17. Blei D. M., Ng A. Y., Jordan M. I. Latent dirichlet allocation. *Journal of Machine Learning Research*. 2003;3:993–1022.
18. Vorontsov K. V., Potapenko A. A. Additivnaya regulyarizatsiya tematicheskikh modeley. *Doklady Akademii nauk*. 2014;456(3):268–271. (In Russ.).
19. Popov O. R. Adaptatsiya mirovykh praktik k probleme dolgosrochnogo tekhnologicheskogo prognozirovaniya sostoyaniya samoorganizuyushchikhsya intellektualnykh system. *Intellektualnye resursy – regionalnomu razvitiyu*. 2021;(2):91–98. (In Russ.).
20. Kramarov S. O., Popov O. R., Dzhariev I. E. et al. Dynamics of link formation in networks structured on the basis of predictive terms. *Russian Technological Journal*. 2023;11(3):17–29. https://doi.org/10.32362/2500-316X-2023-11-3-17-29. (In Russ.).

21. Попов О. Р., Гросу А., Крамаров С. О. Комплексный сетевой алгоритм формирования глоссария контекстно-близких прогностических терминов // Современные информационные технологии и ИТ-образование. 2023. Т. 19, № 3. URL: http://sitito.cs.msu.ru/index.php/SITITO/article/view/999 (дата обращения: 25.03.2024).
22. Chuang J., Gupta S., Manning C. et al. Topic model diagnostics: Assessing domain relevance via topical alignment // International conference on machine learning. 2013. P. 612–620.

Информация об авторах

О. Р. Попов – кандидат технических наук, доцент.

С. О. Крамаров – доктор физико-математических наук, профессор.

21. Popov O. P., Grosu A., Kramarov S. O. Complex network algorithm for glossary formation context-related predictive terms. *Modern Information Technologies and IT-Education*. 2023;19(3). URL: http://sitito.cs.msu.ru/index.php/SITITO/article/view/999 (accessed: 25.03.2024). (In Russ.).
22. Chuang J., Gupta S., Manning C. et al. Topic model diagnostics: Assessing domain relevance via topical alignment. *International conference on machine learning*. 2013:612–620.

About the authors

O. R. Popov – Candidate of Sciences (Engineering), Docent.

S. O. Kramarov – Doctor of Sciences (Physics and Mathematics), Professor.