DOI 10.37882/2223-2966.2025.12.17

КЛЮЧЕВЫЕ ПОДХОДЫ К АВТОМАТИЧЕСКОМУ ИНДЕКСИРОВАНИЮ НАУЧНЫХ ТЕКСТОВ КЛЮЧЕВЫМИ ТЕРМИНАМИ НА ОСНОВЕ ПЕРЕДОВЫХ МЕТОДОВ И МОДЕЛЕЙ ВЕКТОРНЫХ ПРЕДСТАВЛЕНИЙ

KEY APPROACHES TO AUTOMATIC INDEXING OF SCIENTIFIC TEXTS WITH KEY TERMS BASED ON ADVANCED METHODS AND VECTOR REPRESENTATION MODELS

I. Komarov

Summary. This article describes the features of automatic indexing of scientific texts with key terms based on methods and models of vector representations. Within the framework of the analysis, such classes of methods as classical vector representations of words, contextual vector structures, graph-weighted schemes and solutions adapted to specialized corpora are identified, allowing to compare semantic elements of scientific publications and form semantic descriptions of texts. In addition, the following models are considered: pre–trained language structures, contextual transformer architectures, models of vector representation of sentences and documents, as well as generative encoder-decoder solutions. According to the results of the study, four key approaches to automatic indexing of scientific texts with key terms have been identified: embedding-ranking, contextual-transformative with sequence markup, generative approach based on encoder–decoder architecture and large-scale models, structural-embedding hybrid.

Keywords: automatic indexing, keywords, vector representations, methods, models, approaches, scientific texts.

Комаров Иван Дмитриевич
Аспирант, Всероссийский институт научной и технической информации РАН
i.komaroni@ya.ru

Аннотация. В данной статье раскрываются особенности автоматического индексирования научных текстов ключевыми терминами на основе методов и моделей векторных представлений. В рамках анализа выделены такие классы методов, как классические векторные представления слов, контекстные векторные структуры, графово-взвешенные схемы и решения, адаптированные к специализированным корпусам, позволяющие сопоставлять смысловые элементы научных публикаций и формировать семантические описания текстов. Кроме того, рассмотрены следующие модели: предобученные языковые структуры, контекстные трансформерные архитектуры, модели векторного представления предложений и документы, а также генеративные решения типа кодировщик–декодировщик. По результатам исследования выделено четыре ключевых подхода к автоматическому индексированию научных текстов ключевыми терминами: эмбеддинг-ранжировочный, контекстно-трансформерный с разметкой последовательности, генеративный подход на основе архитектуры типа кодировщик–декодировщик и моделей большого масштаба, структурно-эмбеддинговый гибридный.

Ключевые слова: автоматическое индексирование, ключевые термины, векторные представления, методы, модели, подходы, научные тексты.

Введение

Глобальный массив научных текстов за последние десятилетия увеличивается с темпами, которые приближаются к экспоненциальным, что создаёт перегрузку информационных систем и исследовательских коллективов. По данным аналитических отчётов, мировой ежегодный выпуск статей в областях науки и техники вырос с примерно 2,2 млн публикаций в 2014 г. до более 3,3 млн в 2023 г., причём значительная часть прироста связана с авторами из Китая и Индии [15]. Исследования динамики публикационной активности подтверждают сохранение экспоненциальной траектории и средний годовой прирост объёма статей порядка нескольких процентов в год, что усиливает нагрузку на традиционные механизмы рецензирования и индексирования научной информации [14]. Автоматическое индексирование научных текстов ключевыми терминами в таких условиях необходимым элементом инфраструктуры доступа к результатам научных исследований, поскольку ключевые слова задают структуру тематического поиска, аналитических витрин и систем научной оценки. В настоящее время в среднем около трети авторских ключевых слов совпадает с формулировками в заголовке и чуть более половины с формулировками в аннотации [5], что подчёркивает неполную согласованность между формальной разметкой и реальным содержанием публикаций и усиливает значимость автоматического индексирования научных текстов ключевыми терминами.

Актуальность работы

Автоматическое индексирование научных текстов ключевыми терминами опирается на методы извлечения ключевых слов и ключевых фраз, которые за последние

годы претерпели смену доминирующих парадигм от статистических и графовых моделей к методам, основанным на векторных представлениях слов и предобученных языковых моделях [8]. Современные обзорные работы по автоматическому выделению ключевых фраз показывают, что векторные представления функционируют как единый семантический слой, на котором строятся модели на основе классических векторных представлений слов (word embeddings) и архитектуры глубокого обучения, включая трансформеры (transformers) и схемы типа «кодировщик–декодировщик» (encoder–decoder) [12]. Исследовательский интерес также смещается к автоматическому индексированию текстов в рамках предобученных языковых моделей; в таких исследованиях модели векторных представлений (далее — эмбеддинги) используются как для ранжирования кандидатов, так и для формирования новых ключевых терминов в соответствии с контекстом научного текста [10]. Работы, посвящённые глубокому обучению и методам на основе эмбеддингов для извлечения ключевых фраз, дополняют эту картину анализом графовых моделей на эмбеддингах и гибридных схемах, объединяющими статистические признаки с контекстными векторными представлениями и ориентированными на сложные корпуса научных текстов [6]. В совокупности тренды подтверждают необходимость системного рассмотрения современных подходов к автоматическому индексированию научных текстов ключевыми терминами на основе передовых методов и моделей векторных представлений.

Материалы и методы

На первом этапе исследования проводится систематизация современных теоретических и обзорных работ по автоматическому индексированию научных текстов. После этого отобранные материалы группируются по ключевым методам и моделям векторных представлений, затем сопоставляются по единым критериям для выделения ключевых подходов.

Результаты и их обсуждение

В начале целесообразно представить современные методы векторных представлений, используемые в рамках автоматического индексирования научных текстов. Следует отметить, что на практике существуют сотни вариаций автоматического индексирования научных текстов. При этом оно может быть полезно не только для научных текстов, но и для автоматизации и оптимизации инвестиционных сервисов [1], формализации и оптимального управления цифровыми сервисами [2] или балансировки нагрузки в гетерогенной среде вычислительной системы [3]. В связи с чем логично выделить классы методов.

Современные методы векторных представлений, используемые в автоматическом индексировании научных текстов ключевыми терминами, опираются на преобразование лексических единиц и фрагментов текста в точки многомерного пространства, в котором расстояния между объектами отражают семантическую близость. Векторные представления слов (word embeddings) при таком подходе выступают универсальным числовым кодом, который заменяет разреженные частотные описания и снижает размерность признакового пространства [8]. Исторически базовый уровень методов векторных представлений связан с моделями непрерывного кодирования лексики, для которых векторные представления слов (word embeddings) строятся на основе статистики соседства слов в контекстном окне, что реализуется в архитектурах непрерывного мешка слов (CBOW — Continuous Bag of Words) и модели с прогнозированием контекста (Skip-gram). В автоматическом индексировании научных текстов ключевыми терминами данные методы позволяют вычислять степень схожести их векторных описаний между кандидатами в ключевые термины и усреднёнными векторами тематических областей [6].

Дальнейшее развитие методов векторных представлений связано с контекстными векторными представлениями (contextual embeddings), в рамках которых каждая позиция в последовательности получает собственный вектор с учётом окружения, а не фиксированное представление слова. Архитектура трансформеров (transformers) с механизмом самовнимания (self-attention) задаёт основу для построения контекстных представлений, что позволяет учитывать длинные зависимости внутри научных текстов и более точно различать значения терминов в формулировках заголовков, аннотаций и основного текста [9].

Переход к предобученным языковым моделям (pre-trained language models) усилил методическую базу автоматического индексирования научных текстов ключевыми терминами, поскольку векторные представления слов и фраз в таких моделях уже содержат знания, накопленные на больших корпусах. Для вычисления векторных представлений фрагментов текста широко применяются агрегация векторов токенов, использование специального классификационного токена в трансформерных моделях и специализированные процедуры агрегирования предложений для ранжирования кандидатов в ключевые термины по их близости к вектору научного текста [10].

Векторные представления в современных методах индексирования научных текстов ключевыми терминами всё чаще сочетаются с графовыми структурами, в которых вершины соответствуют токенам или фразам, а рёбра отражают семантическую и структурную связанность. В таком случае векторные представления слов (word embeddings) дополняются характеристиками цен-

тральности в графе, а веса векторов модифицируются с учётом структурной значимости позиции токена, что демонстрирует, например, модель с центральным взвешиванием векторных представлений в трансформерной архитектуре для извлечения ключевых фраз из научных текстов [13].

Современные методы векторных представлений в задачах, близких к автоматическому индексированию научных текстов ключевыми терминами, всё более ориентированы на предметно-специализированные корпуса и сложные текстовые структуры, к которым относятся патентные описания и технические документы. Так, предобученные языковые модели (pre-trained language models), адаптированные к патентному корпусу, используют модифицированные процедуры построения векторных представлений, включая комбинирование текстовой и графической информации [7].

Таким образом, можно обобщить рассмотренные методы (табл. 1).

Развитие методов векторных представлений создаёт основу для рассмотрения моделей, использующих их для автоматического индексирования научных текстов ключевыми терминами.

Модели предобученных языковых представлений (pre-trained language models) формируют векторное описание текста на основе глубоких механизмов распределения смысловых связей, что обеспечивает устойчивое извлечение значимых терминологических единиц и поддержку автоматического индексирования научных текстов ключевыми терминами. Внутренняя структура моделей предобученных языковых представлений основана на многоаспектной обработке контекстов, что повышает чувствительность к смысловым границам фраз и их взаимодействию в научных публикациях. Создаваемое подобными моделями векторное описание допускает согласованное сопоставление элементов разных фрагментов текста [11].

Контекстные модели трансформерного типа, использующие архитектуру трансформеров (transformers), обеспечивают многомерное векторное представление последовательностей с учётом распределённых зависимостей, что позволяет выделять ключевые термины в сложных научных корпусах. Механизм самовнимания внутри структуры данной модели усиливает значимые элементы текста и даёт точное различение терминов в разных смысловых окружениях. Важно отметить, что формируемое трансформерными моделями представ-

Таблица 1.

Современные методы векторных представлений, используемые при автоматическом индексировании научных текстов ключевыми терминами

Методы	Краткая характеристика	Роль
Классические векторные представления слов (word embeddings на основе моделей Word2Vec и GloVe)	Плотные векторы фиксированной размерности для слов, получаемые на основе статистики соседства в текстах, с реализацией в архитектурах непрерывного мешка слов (Continuous Bag of Words, CBOW) и модели с прогнозированием контекста (Skip-gram)	Оценка семантической близости кандидатов в ключевые термины к тематике научного текста и сглаживание разреженности частотных признаков при выборе ключевых терминологических единиц
Контекстные векторные представления на основе трансформеров (transformers)	Контекстно-зависимые векторы для каждой позиции в последовательности, формируемые механизмом самовнимания с учётом распределения терминов по всему научному тексту	Уточнение значений ключевых терминов в зависимости от окружения в заголовке, аннотации и основной части текста и повышение точности автоматического индексирования по семантике
Векторные представления фрагментов текста в предобученных языковых моделях (pre-trained language models)	Векторы предложений и документов, получаемые на основе агрегации контекстных векторных представлений слов и специальных классификационных токенов, обученных на крупных корпусах текстов	Сопоставление векторного представления научного текста с векторными представлениями кандидатов в ключевые термины для выбора набора ключевых единиц терминов, согласованного с общей семантикой текста
Графово-взвешенные векторные представления (centrality-weighted embeddings)	Векторные представления слов и фраз в сочетании с характеристиками центральности в графе, построенном на основе семантических и структурных связей внутри научного текста	Усиление вклада ключевых терминов, которые занимают центральное положение в тексте, за счёт модификации векторов и ранжирования при автоматическом индексировании
Методы адаптации векторных представлений к предметно-специализированным корпусам	Векторные представления, адаптированные к терминологии и композиции научно-технических документов, с учётом специфики разделов, ссылок и формализованных описаний	Автоматическое индексирование научных текстов ключевыми терминами в предметно-ориентированных научно-технических трудах

Источник: составлено автором на основе [5–13]

ление поддерживает ранжирование терминологических кандидатов на основе их смысловой связи с основным текстом [9].

Модели, работающие с векторным представлением предложений и документов, используют агрегирование контекстных представлений токенов и специальные классификационные токены, что создаёт целостную векторную структуру для сопоставления длинных научных текстов с кандидатами в ключевые термины. Такие модели могут объединять смысловые связи на уровне отдельных предложений и тематических блоков, что особенно важно для сложных научных текстов. Формируемая структура способствует автоматическому индексированию научных текстов ключевыми терминами в условиях высокой насыщенности контекстами [12].

Модели, поддерживающие графово-взвешенное представление векторов, включают в себя векторные представления слов и фраз в структуру графа и присваивают им веса с учётом центральности, что повышает точность выделения ключевых терминов в научных текстах. Внутренний механизм таких моделей сочетает контекстные представления с характеристиками структурной значимости [13].

Модели, обученные на специализированных научно-технических или патентных корпусах исследований, адаптируют векторное представление к структуре формализованных текстов, что улучшает выделение ключевых терминов в предметных областях. Векторное описание внутри таких моделей отражает специфику терминологии, композицию документа и характер структурных связей между частями текста. Адаптация позволяет повысить качество индексирования текстов ключевыми терминами в узкоспециализированных коллекциях [7].

Наконец, генеративные модели с архитектурой типа кодировщик–декодировщик (encoder–decoder) используют векторное представление текста для формирования новых ключевых терминов, которые отражают смысловые элементы научного описания и расширяют набор терминологических единиц за пределы прямого текстового соответствия. Особенность таких моделей заключается в возможности получать векторные структуры, способные связывать элементы текста с обобщающими формулировками терминов [5].

Таким образом, можно обобщить основные указанные модели (табл. 2).

На основании изученных передовых методов и моделей можно предложить соответствующую концептуальную схему (рис. 1).

Таблица 2.

Современные модели векторных представлений, используемые при автоматическом индексировании научных текстов ключевыми терминами

Модель	Краткая характеристика	Функция модели
Предобученная языковая модель	Многоуровневая архитектура глубокого обучения, формирующая векторные структуры на основе распределённых смысловых связей в больших корпусах	Создание устойчивого векторного описания текста, обеспечивающего согласованное выделение ключевых терминов в насыщенных научных материалах
Контекстная трансформерная модель	Архитектура с использованием механизма самовнимания, задающая контекстно-зависимые векторные представления последовательностей	Различение значений терминов в разных смысловых окружениях и поддержка ранжирования терминологических кандидатов по степени смысловой связи с текстом
Модель векторного представления предложений и документов	Агрегация токенов и специальных классификационных элементов для получения целостного представления длинных текстовых структур	Сопоставление фрагментов научного текста с кандидатами в ключевые термины и повышение точности отбора терминов в условиях сложной структуры документа
Графово-взвешенная модель векторных представлений	Объединение контекстных векторных структур с метрическими характеристиками центральности в графе текстовых элементов	Усиление весов терминов, имеющих ключевое положение в структуре научного текста, и повышение точности последующего ранжирования
Предметно-специализированная предобученная модель	Векторные структуры, адаптированные к терминологии и композиции научно-технических и патентных материалов	Улучшение выделения ключевых терминов в специализированных корпусах и предметных областях
Генеративная модель кодировщик–декодировщик	Архитектура порождения терминов на основе сопоставления векторных описаний текста и обучающих пар «текст–ключевые термины»	Формирование новых терминов, отражающих содержание научного текста и расширяющих итоговое множество ключевых терминов за счёт обобщающих формулировок

Источник: составлено автором на основе [5–13]

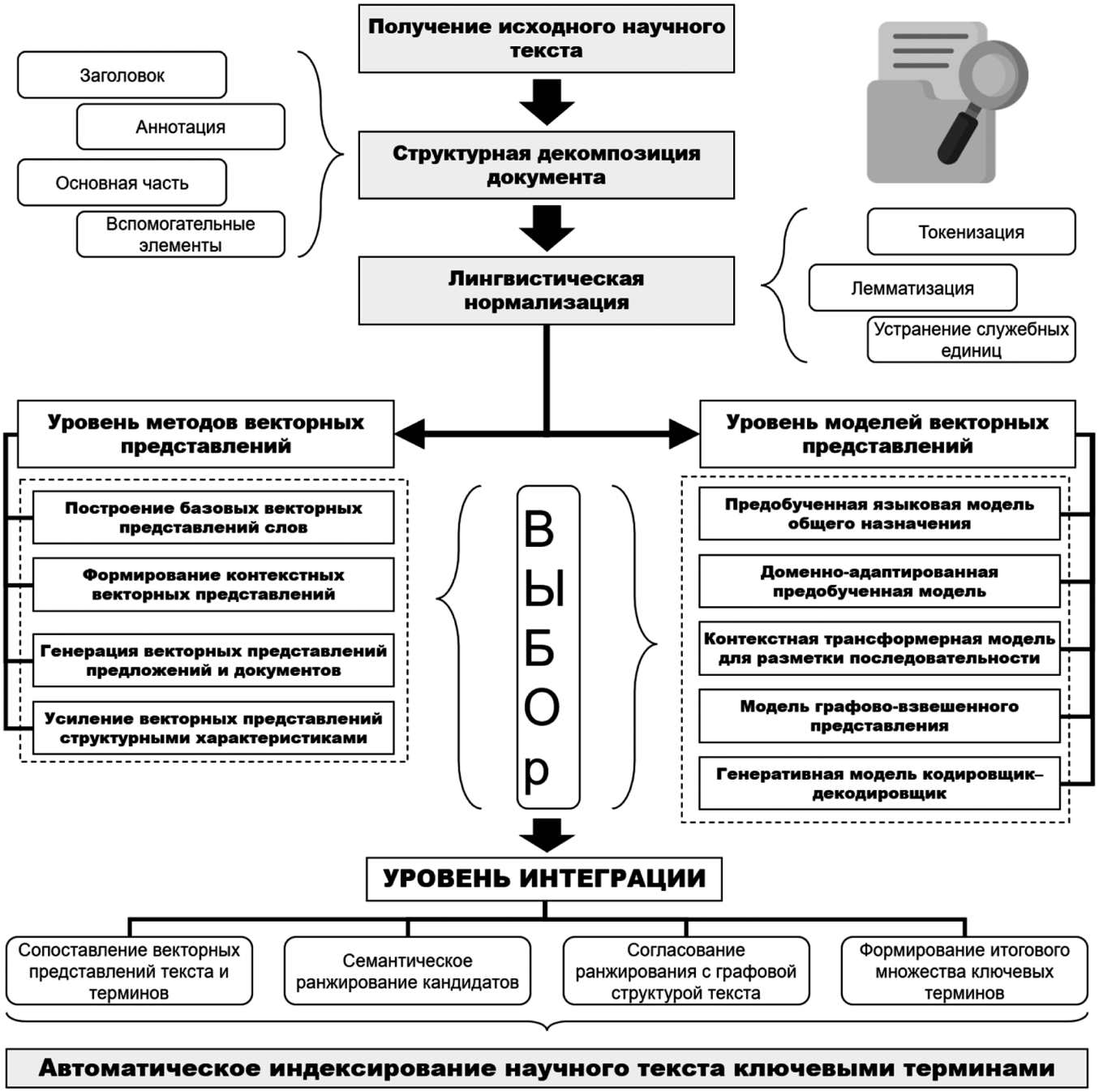

Рис. 1. Концептуальная схема автоматического индексирования научных текстов ключевыми терминами на основе векторных представлений

Источник: составлено автором.

Концептуальная схема отражает переход от структурной обработки научного текста и формирования векторных представлений разного уровня к использованию специализированных моделей, объединяющих контекстные, структурные и генеративные механизмы.

На основании изученных передовых методов и моделей, в целом можно выделить четыре ключевых подхода к автоматическому индексированию научных текстов ключевыми терминами:

— *Эмбеддинг-ранжировочный подход*. В рамках этого подхода кандидатные ключевые термины получают векторные представления слов и фраз, а затем ранжируются по косинусной близости к векторному представлению научного текста. Здесь применяются TF-IDF и векторные представления, модели на основе классических векторных представлений слов и контекстных векторных представлений с последующим ранжированием кандидатов [6], [8], [10], [12].

— *Контекстно-трансформерный подход с разметкой последовательности*. Данный подход предполагает, что модель векторных представлений и выполняет роль классификатора токенов или фраз, и выделяет ключевые термины посредством разметки последовательности с учётом архитектуры трансформеров и предобученных языковых моделей. В данную группу включаются модели наподобие CenBERT-SEQ и более общие PLM-решения для извлечения ключевых фраз [6], [9–10], [13].
— *Генеративный подход на основе архитектуры типа кодировщик–декодировщик и моделей большого масштаба*. В этом подходе векторное представление научного текста подаётся на вход генеративной модели, порождающей набор ключевых терминов, в том числе тех, которые не присутствуют в тексте в явном виде, но отражают его содержание. Сюда относятся универсальные seq2seq-архитектуры и крупные языковые модели, настроенные на автоматическое порождение ключевых терминов для научных статей [5], [10], [12].
— *Структурно-эмбеддинговый гибридный подход*. В рамках данного подхода модели векторных представлений объединяются с информацией о структуре научного текста и графовыми характеристиками, включая центральность элементов, выделенные разделы типа highlights, особенности композиции технических и патентных документов. Сюда входят решения с графово-взвешенными векторными представлениями и предметно-специализированными предобученными моделями [6–7], [11], [13].

Для примера индексирования можно рассмотреть статью Э.С. Манаширова о роли Цезаря в распаде Римского государства [4]. Так, сначала проходит нормализацию заголовка и аннотации, удаление служебных элементов и лемматизацию ключевых фрагментов. На базовой траектории текст представляется матрицей TF-IDF, логистическая регрессия распределяет материал по рубрикам типа «История античности» или «Экономическая история». Далее весь корпус кодируется контекстными описаниями предложений (embeddings) на основе модели BERT, после чего кандидаты упорядочиваются по близости к вектору статьи. Сопоставление двух траекторий показывает, как контекстные представления уточняют набор ключевых терминов и тематическую привязку по сравнению с базовой матрицей TF-IDF.

Заключение

В работе представлены ключевые подходы к автоматическому индексированию научных текстов ключевыми терминами на основе передовых методов и моделей векторных представлений. Итоги проведённого исследования показывают, что автоматическое индексирование научных текстов уже успело превратиться в самостоятельную область развития интеллектуальных инструментов обработки знаний. Рассмотрение современных векторных представлений, контекстных моделей и гибридных архитектур показывает, что повышение качества выделения ключевых терминов связано как с усложнением алгоритмов, так и с их способностью учитывать многоаспектность научного дискурса, вариативность формулировок и скрытую структуру текста. Сформулированные в работе подходы и концептуальная схема отражают движение научного сообщества к созданию гибких и адаптивных систем аналитики, способных работать с быстро растущими объёмами научных публикаций и обеспечивать более точное и устойчивое тематическое представление исследовательских материалов.

ЛИТЕРАТУРА

1. Куровский С.В., Мишин Д.А., Маринин А.К., Бурдик В., Куровская М.А. Современные подходы к автоматизации и оптимизации инвестиционных сервисов телекоммуникационных компаний // Инновации и инвестиции. — 2024. — № 10. — С. 455–460.
2. Куровский С.В., Мишин Д.А., Штыков Р.А. Задачи и методы формализации и оптимального управления цифровыми сервисами в компаниях // Современная наука: актуальные проблемы теории и практики. Серия: Естественные и технические науки. — 2024. — № 10–2. — С. 39–45.
3. Куровский С.В., Мишин Д.А., Шульман В.Д. Алгоритм балансировки нагрузки в гетерогенной среде вычислительной системы // Международный научно-исследовательский журнал. — 2025. — № 5 (155). — С. 1–10.
4. Манаширов Э.С. Тени цезаря: историко-экономический анализ негативного влияния на римскую экономику // Инновации и инвестиции. — 2025. — № 10. — С.92–96.
5. Asar M.A., Mitincik A., Turhan S., Orman G.K. Automated Keyword Generation for Academic Research Articles Using Large Language Models // Procedia Computer Science. — 2025. — Vol. 270. — P. 2215–2224.
6. Giarelis N., Karacapilidis N. Deep Learning and Embeddings-Based Approaches for Keyphrase Extraction: A Literature Review // Knowledge and Information Systems. — 2024. — Vol. 66. — P. 6493–6526.
7. Jiang L., Goetz S.M. Natural Language Processing in the Patent Domain: A Survey // Artificial Intelligence Review. — 2025. — Vol. 58, No. 214. — P. 1–62.
8. Nomoto T. Keyword Extraction: A Modern Perspective // SN Computer Science. — 2023. — Vol. 4, No. 92. — P. 1–19.
9. Song M., Feng Y., Jing L. A Survey on Recent Advances in Keyphrase Extraction from Pre-trained Language Models // Findings of the Association for Computational Linguistics: EACL 2023. — 2023. — P. 2153–2164.

10. Umair M., Sultana T., Lee Y.-K. Pre-trained Language Models for Keyphrase Prediction: A Review // ICT Express. — 2024. — Vol. 10. — P. 871–890.
11. Xiang Y., Yan X., Zhang C. Enhancing Keyword Extraction from Academic Articles Using Highlights // ASIS&T Annual Meeting Proceedings. — 2024. — P. 1147–1150.
12. Xie B., Song J., Shao L., Wu S., Wei X., Yang B., Lin H., Xie J., Su J. From Statistical Methods to Deep Learning, Automatic Keyphrase Prediction: A Survey // Information Processing and Management. — 2023. — Vol. 60, No. 4. — P. 1–21.
13. Zengeya T., Fonou Dombeu J.V., Gwetu M. A Centrality-Weighted Bidirectional Encoder Representation from Transformers Model for Enhanced Sequence Labeling in Key Phrase Extraction from Scientific Texts // Big Data and Cognitive Computing. — 2024. — Vol. 8, No. 182. — P. 1–33.
14. Excessive growth in the number of scientific publications // Ouvrir la science. Available online: https://www.ouvrirlascience.fr/excessive-growth-in-the-number-of-scientific-publications/ (accessed on 24.11.2025).
15. Publication Output by Geography and Scientific Field. Discovery: R&D Activity and Research Publications. Available online: (accessed on 24.11.2025).

Журнал «Современная наука: актуальные проблемы теории и практики»