Russian Technological Journal.
2026;14(3):24–42

ISSN 2782-3210 (Print)
ISSN 2500-316X (Online)

УДК 004.912
https://doi.org/10.32362/2500-316X-2026-14-3-24-42
EDN BMHCUK

НАУЧНАЯ СТАТЬЯ

Разработка прикладных инструментов установления информационного морфизма при анализе текстовых документов на основе семантико-онтологической и графовой моделей

Н.С. Курдюков [@], В.Н. Калинин, С.А. Кудж, Д.О. Жуков

МИРЭА – Российский технологический университет, Москва, 119454 Россия
[@] Автор для переписки, e-mail: nskurdyukov@gmail.com

• **Поступила:** 19.01.2026 • **Доработана:** 06.02.2026 • **Принята к опубликованию:** 27.03.2026

Резюме

Цели. Исследуется возможность использования семантико-онтологической модели анализа текстовых документов для разработки прикладных инструментов установления информационного морфизма. В качестве текстового онтологического основания для количественного анализа научных текстов рассматриваются паспорта научных специальностей ВАК[1]. Цель работы состоит в разработке графовой семантико-онтологической модели, которая по тексту статьи или автореферата восстанавливает профиль близости к шифрам специальностей и тем самым задает отображение от пространства документов к пространству паспортов.

Методы. Паспорта научных специальностей обрабатываются как единый корпус. По чанкам строится словарь униграмм и биграмм, рассчитываются TF-IDF[2] представления и локальные графы ICAN[3]. Для пар «документ и паспорт» вычисляются меры сходства, которые в лексическом и семантическом слоях сворачиваются в оценки и объединяются в гибридную метрику. Результат переводится в вероятностное распределение по шифрам через температурный softmax[4]. Качество модели оценивается на корпусе

[1] Высшая аттестационная комиссия при Министерстве науки и высшего образования Российской Федерации. https://vak.gisnauka.ru/. Дата обращения 04.04.2026. [Higher Attestation Commission under the Ministry of Science and Higher Education of the Russian Federation. https://vak.gisnauka.ru/. Accessed April 04, 2026. (In Russ.).]

[2] Term frequency – inverse document frequency – статистическая мера, учитывающая частоту термина в документе и обратную частоту документа. [Term frequency – inverse document frequency (TF-IDF) is a statistical measure that takes into account the term frequency in a document and the inverse document frequency.]

[3] Incremental construction of an associative network – вычислительная модель инкрементального построения ассоциативной сети на основе корпуса текстов. [Incremental construction of an associative network (ICAN) is a computational model for incremental construction of an associative network based on a text corpus.]

[4] Температурный softmax – функция нормализации, переводящая логиты z_i в распределение вероятностей, где параметр температуры $T > 0$ регулирует «резкость» этого распределения.

авторефератов и статей из журналов Перечня ВАК РФ[5], дополнительно проводится сравнение с крупными языковыми моделями.

Результаты. Гибридная схема дает точность top 1 около 0.69 и top 3 около 0.90 на авторефератах, а на статьях достигает 0.91 и 0.93. Это выше, чем у лексических и семантических вариантов. Метод выигрывает по top 1 для статей и остается сопоставимым по top 3, сохраняя интерпретируемость через *n*-граммы и контекстные графы.

Выводы. Паспорта ВАК могут быть практичным онтологическим основанием для анализа научных текстов, а предложенная модель является интерпретируемой и вычислительно экономичной альтернативой для выбора шифра и построения тематических профилей с учетом междисциплинарности.

Ключевые слова: паспорта специальностей ВАК, графовая семантико-онтологическая модель, TF-IDF, модель ICAN, информационный морфизм, классификация научных текстов

Для цитирования: Курдюков Н.С., Калинин В.Н., Кудж С.А., Жуков Д.О. Разработка прикладных инструментов установления информационного морфизма при анализе текстовых документов на основе семантико-онтологической и графовой моделей. *Russian Technological Journal*. 2026;14(3):24–42. https://doi.org/10.32362/2500-316X-2026-14-3-24-42, https://www.elibrary.ru/BMHCUK

Прозрачность финансовой деятельности: Авторы не имеют финансовой заинтересованности в представленных материалах или методах.

Авторы заявляют об отсутствии конфликта интересов.

RESEARCH ARTICLE

Development of applied tools for establishing information morphism in the analysis of text documents based on semantic-ontological and graph models

Nikita S. Kurdyukov [@], Vladimir N. Kalinin, Stanislav A. Kudzh, Dmitry O. Zhukov

MIREA – Russian Technological University, Moscow, 119454 Russia
[@] Corresponding author, e-mail: nskurdyukov@gmail.com

• **Submitted:** 19.01.2026 • **Revised:** 06.02.2026 • **Accepted:** 27.03.2026

Abstract

Objectives. The work considers whether a semantic-ontological model for scientific text analysis can support practical tools for establishing information morphism. Using VAK[6] specialty passports as the textual ontological basis, we propose a graph-based model that reconstructs a proximity profile to specialty codes from an article or dissertation abstract to map the document space to the passport space.

[5] Перечень рецензируемых научных изданий, в которых должны быть опубликованы основные научные результаты диссертаций на соискание ученой степени кандидата наук, на соискание ученой степени доктора наук. https://vak.gisnauka.ru/documents/editions. Дата обращения 04.04.2026. [List of peer-reviewed scientific publications in which the main scientific results of dissertations for the degree of candidate of sciences and for the degree of doctor of sciences should be published. https://vak.gisnauka.ru/documents/editions. Accessed April 04, 2026. (In Russ.).]

[6] Higher Attestation Commission under the Ministry of Science and Higher Education of the Russian Federation. https://vak.gisnauka.ru/. Accessed April 04, 2026. (In Russ.).

Methods. Processing the passports as a single corpus, a shared unigram and bigram vocabulary is constructed from their chunks. Term frequency is computed in the form of inverse document frequency (TF-IDF) representations to construct local semantic graphs on the basis of incremental construction of an associative network (ICAN). For each document passport pair, similarity measures are merged into a hybrid metric by aggregation within lexical and semantic layers. Scores are converted into a probability distribution via codes based on temperature softmax functions. The model is evaluated on a corpus of dissertation abstracts and a corpus of articles of VAK list journals[7], and the results are compared with large language models.
Results. The hybrid scheme, which achieves average top 1 accuracy of about 0.69 and top 3 of about 0.90 on abstracts, reaches 0.91 and 0.93 on articles to outperform lexical-only and semantic-only variants. Considered relative to large language models, the hybrid scheme achieves superior top 1 accuracy for articles and comparable accuracy in top 3, while remaining interpretable through n grams and contextual passport graphs.
Conclusions. The proposed model, which uses VAK passports to provide a practical ontological foundation, represents an interpretable and computationally efficient alternative for code selection and thematic profiling that accounts for interdisciplinarity.

Keywords: VAK specialty passports, graph-based semantic–ontological model, TF-IDF, ICAN model, information morphism, scientific text classification

For citation: Kurdyukov N.S., Kalinin V.N., Kudzh S.A., Zhukov D.O. Development of applied tools for establishing information morphism in the analysis of text documents based on semantic-ontological and graph models. *Russian Technological Journal.* 2026;14(3):24–42. https://doi.org/10.32362/2500-316X-2026-14-3-24-42, https://www.elibrary.ru/BMHCUK

Financial disclosure: The authors have no financial or proprietary interest in any material or method mentioned.

The authors declare no conflicts of interest.

ВВЕДЕНИЕ

Классические методы классификации текстов, основанные на модели «мешка слов» и простых векторных представлениях, дают вполне приемлемые результаты в условиях строгой терминологической стандартизации, однако они слабо учитывают синонимию, вариативность формулировок и контекстные связи между терминами.

Современные нейросетевые и графовые модели обработки текстов частично снимают эти ограничения, однако такие модели обычно обучаются на общих корпусах и слабо привязаны к конкретным нормативным онтологиям, что снижает интерпретируемость результата, например, при решении задач, связанных с обработкой официальных документов, основанных на стандартизации, что будет показано в статье далее.

Например, системы аттестации научных кадров и экспертизы публикаций опираются на регламенты тематического отнесения работ к научным специальностям. В российской практике такую роль выполняют паспорта специальностей Высшей аттестационной комиссии (ВАК)[8], которые описывают объекты исследования, типовые задачи и методы и тем самым задают нормативное тематическое пространство. Быстрый рост объема цифровых научных текстов усиливает потребность в автоматизированных методах сопоставления статей и диссертаций с этими нормативными описаниями. Такие методы необходимы для выбора шифров диссертаций и статей, для анализа тематического профиля диссертационных советов и журналов и для мониторинга структуры научной активности в целом.

Одним из решений подобного рода задач может стать разработка точных, но технологичных и быстродействующих инструментов, основанных на использовании онтологических методов и подходов, ставших уже классическими.

В представленной работе как один из возможных примеров реализации онтологического подхода для решения практических задач текстовой аналитики использованы паспорта специальностей ВАК, которые рассматриваются как онтологическое основание для создания необходимых инструментов анализа. На их основе формируется единое признаковое пространство, в котором описываются как сами паспорта, так и научные тексты. Предлагается графовая семантико-онтологическая модель признаков, включающая лексический слой на основе TF-IDF[9] и семантический слой на основе локальных контекстных

[7] List of peer-reviewed scientific publications in which the main scientific results of dissertations for the degree of candidate of sciences and for the degree of doctor of sciences should be published. https://vak.gisnauka.ru/documents/editions. Accessed April 04, 2026. (In Russ.).

[8] Высшая аттестационная комиссия при Министерстве науки и высшего образования Российской Федерации. https://vak.gisnauka.ru/. Дата обращения 04.04.2026. [Higher Attestation Commission under the Ministry of Science and Higher Education of the Russian Federation. https://vak.gisnauka.ru/. Accessed April 04, 2026. (In Russ.).]

[9] Term frequency – inverse document frequency – статистическая мера, учитывающая частоту термина в документе и обратную частоту документа. [Term frequency – inverse document frequency (TF-IDF) is a statistical measure that takes into account the term frequency in a document and the inverse document frequency.]

графов ICAN[10]. Для каждого документа и каждого паспорта определяются локальные и глобальные метрики близости, которые объединяются в гибридную скалярную оценку. Далее эта оценка переводится в вероятностный профиль по шифрам, который интерпретируется как информационный морфизм – структурно-сохраняющее отображение, переводящее представление документа из признакового пространства в онтологическое пространство паспортов таким образом, что лексические и контекстно-семантические связи получают согласованное представление в терминах нормативных понятий специальностей.

Такое построение решает две задачи. С одной стороны, оно сохраняет строгую привязку к нормативному языку паспортов и обеспечивает прозрачную интерпретацию вклада отдельных терминов и контекстных связей. С другой стороны, использование графового семантического слоя и гибридной агрегации позволяет учитывать вариативность формулировок, аббревиатуры и междисциплинарные связи. Это делает возможным не только точечный выбор шифра, но и анализ распределения тематического вклада документа между несколькими близкими специальностями и построение агрегированных профилей научных акторов.

Цель статьи состоит в разработке и экспериментальной оценке графовой семантико-онтологической модели анализа научных текстов на русском языке, основанной в качестве объекта и предмета исследований на паспортах специальностей ВАК, а также в сопоставлении ее с крупными языковыми моделями в задачах восстановления шифров и анализа тематических профилей. Для достижения этой цели формируется корпус паспортов и сопутствующих текстов, разрабатывается единый конвейер нормализации, задаются лексический и семантический слои признаков и интегральная метрика информационного морфизма, а затем проводится валидация на авторефератах соискателей и на статьях из журналов Перечня ВАК[11].

Статья организована следующим образом. В первом разделе приводится обзор работ по семантической классификации текстов, онтологически ориентированным моделям и графовым представлениям. Во втором разделе описываются онтологическое основание исследования и корпус данных. В третьем разделе вводится графовая семантико-онтологическая модель признаков на основе TF-IDF и ICAN. В четвертом разделе формализуются метрики информационного морфизма и вероятностная интерпретация результатов. В пятом разделе представлены результаты валидации на авторефератах и научных статьях и сравнение с крупными языковыми моделями. В шестом разделе обсуждаются полученные результаты и формулируются выводы и направления дальнейших исследований.

1. ЛИТЕРАТУРНЫЙ ОБЗОР

В обзоре [1] систематизированы подходы к семантической классификации текстов и проведено их сопоставление с традиционными методами на основе метода «мешок слов». Авторы показывают ограничения векторного представления BoW[12] (высокая размерность, разреженность, игнорирование синонимии и полисемии) и выделяют пять основных классов семантических методов: основанные на онтологиях, корпусно-статистические, модели глубокого обучения, методы с учетом последовательностей слов/символов и лингвистически обогащенные подходы. Обзор обобщает результаты множества экспериментов и демонстрирует преимущество семантических моделей по точности классификации.

В работе [2] анализируются методы извлечения знаний в контексте семантического веба с акцентом на онтологически ориентированный отбор признаков. Онтологии используются для представления и выбора признаков в задачах классификации, что позволяет сокращать размерность и повышать интерпретируемость моделей, при этом качество подхода существенно зависит от полноты и согласованности онтологий.

В работах [3–5] онтологии и семантические технологии рассматриваются как ключевой инструмент структурирования инженерных и научных знаний. В [3] предлагается онтологически ориентированный подход к автоматической классификации инженерных стандартов, использующий доменные онтологии для сопоставления фрагментов документов с бизнес-подразделениями.

В [4] разрабатывается двухэтапный метод описания заранее определенных потоков информации в стандартах и директивах, включающий

[10] Incremental construction of an associative network – вычислительная модель инкрементального построения ассоциативной сети на основе корпуса текстов. [Incremental construction of an associative network (ICAN) is a computational model for incremental construction of an associative network based on a text corpus.]

[11] Перечень рецензируемых научных изданий, в которых должны быть опубликованы основные научные результаты диссертаций на соискание ученой степени кандидата наук, на соискание ученой степени доктора наук. https://vak.gisnauka.ru/documents/editions. Дата обращения 04.04.2026. [List of peer-reviewed scientific publications in which the main scientific results of dissertations for the degree of candidate of sciences and for the degree of doctor of sciences should be published. https://vak.gisnauka.ru/documents/editions. Accessed April 04, 2026. (In Russ.).]

[12] Bag of words.

обобщенную модель данных и машиночитаемое представление, что повышает доступность и переиспользуемость сведений в цифровых моделях изделий. Работа [5] содержит библиометрический и семантический анализ исследований по онтологиям и семантическому вебу, выделяя основные направления, связанные с динамическим обновлением онтологий и масштабируемостью в условиях Big Data[13].

В работе [6] представлен обзор онтологически ориентированных методов классификации текстов, в котором онтологии рассматриваются как формальная основа для обогащения векторных представлений и архитектур глубокого обучения за счет явного моделирования доменных концептов и их связей. В исследовании [7] предложен онтологически ориентированный подход к извлечению контекстуализированной информации из научных публикаций. На основе трансформерных моделей реализуется двухэтапный конвейер, включающий классификацию предложений и распознавание сущностей (исследовательских действий и методов), а затем извлечение связей между действиями и используемыми методами с интеграцией полученных данных и метаданных публикаций в виде RDF[14]-графа.

Ряд работ [8–10] демонстрирует потенциал лексико-семантических представлений текста для решения разных задач анализа. В [8] вводятся специальные лексико-семантические параметры для определения тональности, получаемые через семантическое расширение сентиментных лексиконов и распределенные представления, что позволяет унифицировать размерность данных и повысить качество классификации.

Авторы работы [9] предлагают модель «лексико-семантических связей» между документами и статистический алгоритм кластеризации на основе косинусной близости в лексико-семантическом пространстве, сопоставимый по качеству с методом Affinity Propagation[15].

В работе [10] используется набор лексико-семантических паттернов и процедуры сопоставления упоминаний в тексте с объектами доменной онтологии для их полуавтоматического пополнения. В совокупности эти подходы подтверждают эффективность лексико-семантических признаков.

В работе [11] исследователи решают задачу многоклассовой классификации несбалансированных текстовых данных с использованием лексической онтологии WordNet[16] и модели BERT[17]. Лексическая онтология WordNet применяется для онтологически ориентированного снижения размерности признаков, после чего выполняется классификация традиционными алгоритмами и предобученной моделью BERT. Эксперименты на семиклассовом корпусе показывают, что сочетание WordNet и BERT обеспечивает наивысшую точность (до 93.77%) по сравнению с вариантами без онтологического уточнения признаков.

Ряд работ [12–15] систематически исследует графовые представления текста и графовые нейросетевые модели для задач классификации. В [12] проводится масштабное сравнение различных схем построения графов и архитектур GNN[18] с трансформерными моделями. В [13] исследователи анализируют, как выбор стратегии представления текста в виде графа влияет на качество GNN-моделей. Работы [14] и [15] развивают использование графовых моделей в архитектурах анализа текстового контента, ориентируясь на повышение интерпретируемости и качества классификации. В [14] предложен фреймворк, в котором графовое представление текста дополняется символическим отбором признаков, а в [15] разрабатывается иерархический графовый фреймворк, объединяющий лингвистические признаки, доменную онтологию, многоуровневое обучение GNN, механизмы внимания и динамическое слияние с BERT.

В работе [16] предложена концепция информационного онтологического моделирования, в которой информационный поиск и методы кластеризации рассматриваются как стадии формирования онтологических моделей и извлечения неявного знания.

В работе [17] авторы исследуют, как топологические характеристики графа слов зависят от жанра текста. Тексты разных жанров представляются в виде графов, где вершины соответствуют словам, а ребра – их соседству в биграммах; показано, что параметры таких сетей систематически различаются между жанрами и могут использоваться для

[13] Большие данные.

[14] Resource description framework – среда описания ресурса – модель представления данных.

[15] Алгоритм кластеризации в машинном обучении, основанный на концепции «обмена сообщениями» между точками данных. [A clustering algorithm in machine learning based on the concept of messaging between data points.]

[16] Лингвистическая онтология, электронный тезаурус, который представляет в виде иерархической структуры систему значений слов общезначимого английского языка. [Linguistic ontology, an electronic thesaurus that represents the system of meanings of words in the generally significant English language in the form of a hierarchical structure.]

[17] Bidirectional encoder representations from transformers – нейронная сеть, разработанная для улучшения понимания естественного языка. Главное отличие BERT от предыдущих моделей – двунаправленное понимание контекста. [Bidirectional encoder representation from transformers (BERT) is a neural network designed to improve natural language understanding. BERT's key difference from previous models is its bidirectional context understanding.]

[18] Graph neural network – графовая нейронная сеть.

их характеристики и выявления поджанров. В работе [18] используется модель ICAN для построения семантических графов на уровне отдельных документов, что позволяет учитывать порядок слов, важный для семантического анализа.

2. ОНТОЛОГИЧЕСКОЕ ОСНОВАНИЕ И КОРПУС ДАННЫХ

2.1. Паспорта специальностей ВАК

В системе аттестации научных кадров РФ тексты паспортов научных специальностей ВАК выполняют функцию нормативного описания предметных областей. Каждый паспорт закреплен за определенным шифром специальности и задает совокупность типичных объектов исследования, классов задач, применяемых методов и областей использования результатов. Эти документы имеют надведомственный статус и служат формальным основанием для отнесения диссертаций, авторефератов и публикаций к конкретной научной области.

Содержательно паспорта обладают устойчивой внутренней структурой, в которой, как правило, выделяются следующие блоки:

- перечень основных направлений и объектов исследований;
- типовые научные задачи и методы;
- связанные области знаний.

Такая структура позволяет использовать паспорта специальности не только как регламентирующие документы, но и как текстовые представления онтологий предметных областей. Под онтологией здесь понимается явным образом заданная система понятий и отношений, ограничивающая допустимое тематическое пространство.

Множество паспортов специальностей можно трактовать как конечное множество онтологических объектов:

$$O = \{s_1, \ldots, s_N\}, \quad (1)$$

где каждый элемент s_i соответствует одной научной специальности, а текст паспорта задает ее концептуальное содержание.

Иерархическая структура шифров (укрупненные области, группы, конкретные специальности) задает частичный порядок на множестве O и фиксирует отношения включения и близости между областями. Вводится онтологическое пространство научных специальностей, в котором уровень укрупненных областей соответствует более общим классам, а отдельные шифры задают специализированные поддомены.

Кроме того, принимается ключевое методологическое допущение. Текст паспорта рассматривается как прототип соответствующей предметной области. Лексемы и устойчивые словосочетания интерпретируются как поверхностные реализации понятий и отношений данной области, а структура разделов отражает тематическую связность этих элементов. Это позволяет использовать корпус паспортов научных специальностей для построения опорного семантико-онтологического признакового пространства, в котором описываются как сами паспорта, так и анализируемые документы.

При этом паспорта не являются формальными онтологиями в строгом логическом смысле и не задают систему аксиом и строгих ограничений. Однако их официальный статус, стабильность структуры и ориентация на описания объектов, задач и методов позволяют рассматривать их как практически приемлемое онтологическое основание для информационного анализа научных текстов.

В дальнейшем именно это основание используется для построения семантико-онтологической модели признаков и определения информационного морфизма между документами и пространством научных специальностей.

2.2. Сбор и хранение текстовых данных

Корпус онтологического основания формируется из паспортов научных специальностей, утвержденных ВАК. Для их сбора используются специализированные программные парсеры, которые автоматически обходят связанные веб-страницы официальных ресурсов, загружают файлы паспортов и привязанные к ним документы (в форматах html и pdf), а также фиксируют технические метаданные.

Для документов в формате pdf выполняется проверка наличия текстового слоя. Если текстовый слой присутствует, извлечение осуществляется напрямую; в противном случае применяется оптическое распознавание текста (optical character recognition, OCR) с последующей базовой очисткой результатов.

На этом этапе текст еще не нормализуется лингвистически, но устраняются очевидные артефакты (случайные бинарные вставки, некорректные кодировки), чтобы обеспечить корректность дальнейшей обработки. Подробный алгоритм нормализации описывается далее.

Для каждого паспорта сохраняются, в частности, следующие поля:

- уникальный идентификатор записи,
- код специальности,
- наименование,
- URL[19]-первоисточник,
- время сбора,
- хеш-значение файла,
- извлеченный текст после первичной очистки.

[19] Uniform resource locator – единый указатель ресурсов.

Первичный корпус паспортов формируется в виде JSONL[20]-файлов (один объект на строку), что удобно для архивирования и обмена данными. Для долгосрочного хранения и эффективного доступа данные переносятся в реляционную систему управления базами данных PostgreSQL[21].

Для паспортов специальностей создается отдельная таблица, включающая код и наименование специальности, метаданные источника, контрольную сумму файла, а также нормализованный текст паспорта, используемый при построении признаков.

Для входных документов (научных статей и авторефератов диссертаций) создается соответствующая таблица, в которой, помимо текстового содержимого и типа документа, при наличии фиксируется исходный код специальности (для авторефератов), т.к. он используется в качестве экспертной оценки при оценке качества модели.

2.3. Нормализация текстов

Нормализация текстов играет ключевую роль в предлагаемой модели, поскольку именно на этом этапе задается единое признаковое пространство, в котором в дальнейшем будут сравниваться научные тексты с паспортами научных специальностей.

Обработка текстового контента намеренно организована как общая для всех категорий документов, чтобы различия в признаках отражали содержательные особенности текстов, а не артефакты формата, верстки или источника.

На первом этапе выполняется техническая очистка извлеченного текста. После получения содержимого из html-страниц или pdf-файлов (с использованием OCR при отсутствии текстового слоя) устраняются некорректные кодировки, бинарные вставки, фрагменты служебной разметки. После переносы строк приводятся к обычным пробелам. Для pdf-документов автоматически выявляются и удаляются повторяющиеся колонтитулы и шапки страниц.

Затем выполняется фильтрация служебных блоков. Из текстов удаляются явно редакционно-издательские фрагменты (литература, ключевые слова, сведения об издательстве, контактная информация авторов, сведения об авторах и прочее), а также технические элементы: URL, адреса электронной почты, числовые идентификаторы (ORCID[22], Scopus ID[23] и др.).

Следующим шагом является приведение форм текста. Весь текст переводится в нижний регистр, символ «ё» заменяется на «е», устраняются искусственные переносы и убираются лишние пробелы. Удаляется пунктуация и все служебные символы. Отдельное внимание уделяется дефисным терминам, т.к. для сохранения терминологической целостности используется словарь склеек, в соответствии с которым сложные слова вида «объектно-ориентированный», «научно-исследовательский» приводятся к форме с нижним подчеркиванием («объектно_ориентированный», «научно_исследовательский»). Это позволяет рассматривать устойчивые сложные термины как единые лексические единицы в последующем анализе.

Далее проводится обработка числовых выражений. Если токен состоит только из цифр, он преобразуется в словесную форму на русском языке (например, «2025» преобразуется в «две тысячи двадцать пятый»), что предотвращает увеличение размерности словаря за счет большого количества уникальных чисел при сохранении смысловой информации о количественных характеристиках. Однако при этом технические стандарты, индексы и обозначения (например, «5g», «3d», «h264», «ipv6») сохраняются как отдельные токены и нормализуются к нижнему регистру, поскольку они являются важными маркерами предметной области.

На этапе лингвистической нормализации выполняются токенизация, лемматизация и фильтрация стоп-слов. Текст разбивается на токены с учетом ранее введенных правил замены пунктуации и дефисных конструкций. Для каждого токена определяется нормальная форма (лемма). При необходимости унифицируются распространенные сокращения и обозначения единиц измерения, а доменные аббревиатуры переводятся в стандартные формы.

В итоге результатом нормализации является устойчивое текстовое представление в виде последовательности лемм, разделенных пробелами, без пунктуации, чисел и служебных слов. При необходимости дополнительно формируется список токенов с позиционными диапазонами в исходном тексте, что позволяет позже интерпретировать вклад отдельных фрагментов в итоговые оценки близости. Именно эта нормализованная последовательность лемм используется для построения признаков в лексическом (TF-IDF) и семантическом (ICAN) слоях модели.

[20] JSON Lines – текстовый формат, в котором каждая строка содержит один допустимый JSON-объект (JavaScript object notation). [JSON Lines is a text format in which each line contains one valid JSON object (JavaScript object notation).]

[21] https://www.postgresql.org/. Дата обращения 04.04.2026. / Accessed April 04, 2026.

[22] Open researcher and contributor identifier – открытый идентификатор исследователя и участника.

[23] Уникальный идентификатор автора в базе данных Scopus. https://www.scopus.com/. Дата обращения 04.04.2026. [Unique author identifier in the Scopus database. https://www.scopus.com/. Accessed April 04, 2026.]

3. ГРАФОВАЯ СЕМАНТИКО-ОНТОЛОГИЧЕСКАЯ МОДЕЛЬ ПРИЗНАКОВ

3.1. «Чанки» паспортов и *n*-граммы

Нормализованные тексты паспортов научных специальностей (см. 2.3) далее рассматриваются как последовательности лемм w:

$$d_s = (w_1, ..., w_{T_s}), \quad (2)$$

где T_s – длина паспорта s в токенах после предобработки.

Поскольку паспорта специальностей состоят из разнородных разделов (описание объектов, задач, методов, междисциплинарных связей), то представление всего текста одним вектором приводит к тому, что слабосвязанные фрагменты усредняются, что снижает чувствительность модели к тематическим различиям внутри паспорта.

Для повышения чувствительности модели к локальным областям смысловой концентрации каждый паспорт разбивается на фрагменты фиксированной длины L токенов – чанки. Число чанков для паспорта s задается следующим образом:

$$n_s = \left\lceil \frac{T_s}{L} \right\rceil. \quad (3)$$

Каждый чанк $c_{s,i}$ наследует метаданные паспорта (код и наименование специальности), а также хранит позиционный диапазон $[a_i, b_i]$ в исходном тексте, что позволяет впоследствии интерпретировать вклад отдельных фрагментов в итоговые оценки.

Выбор длины L определяется компромиссом между локальной чувствительностью и устойчивостью представления.

Слишком короткие чанки приводят к чрезмерной разреженности векторного пространства и усиливают влияние артефактов разметки и ошибок OCR на каждый отдельный фрагмент. Слишком длинные, напротив, делают локальные метрики сходства практически эквивалентными глобальной, что фактически нивелирует эффект разбиения текста на чанки.

В проведенных экспериментах анализировались значения $L \in \{100, 200, 300\}$, длина $L = 200$ показала баланс между числом фрагментов на паспорт, устойчивостью к шуму и вычислительной стоимостью и далее используется как значение по умолчанию (рисунок).

Разбиение на чанки решает две задачи. Во-первых, локализует профильную лексику, а именно: внутри фрагмента длиной порядка сотен токенов увеличивается доля терминов, характерных для конкретной области, что приводит к более «резким» косинусным сходствам при сравнении с документами.

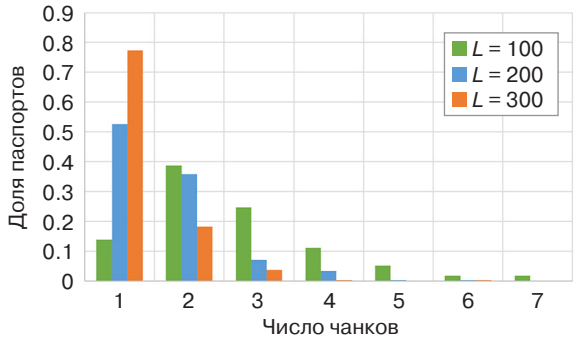

Рисунок. Распределение числа чанков n_s по паспортам при $L \in \{100, 200, 300\}$

Во-вторых, повышает устойчивость к структурным и версточным артефактам, в которых отдельные технические вставки или остатки колонтитулов влияют лишь на свои локальные вектора и не доминируют в глобальной оценке.

Поэтому наличие у значимой доли паспортов как минимум двух-трех фрагментов важно для локальной метрики, т.к. увеличивается вероятность того, что хотя бы один фрагмент будет близок к рассматриваемому документу, а агрегат по top N становится чувствительнее к тематическим совпадениям.

После на множестве всех чанков паспортов $D = \{c_1, \ldots, c_N\}$ строится лексический словарь n-грамм. Для каждого чанка $c \in D$ имеется последовательность лемм $(w_1, \ldots, w_{T(c)})$.

Рассматриваются два уровня:

- множество униграмм $U(c) = \{w_i\}_{i=1}^{T(c)}$,
- множество биграмм $B(c) = \{(w_i, w_{i+1})\}_{i=1}^{T(c)-1}$.

Использование биграмм принципиально для фиксации устойчивых терминологических сочетаний («нейронная сеть», «интернет-сеть», «сеть связи» и др.), которые уточняют семантику полисемичных лемм и служат более надежными маркерами предметных областей, чем отдельные слова.

Для каждого термина t (как униграммы, так и биграммы) вычисляется число чанков корпуса, в которых этот термин встречается хотя бы один раз:

$$df(t) = |\{c \in D: t \in c\}. \quad (4)$$

Далее вводятся пороговые ограничения по минимальному и максимальному числу появлений:

$$df(t) \geq d_{MIN}, \frac{df(t)}{N} \leq d_{MAX}, \quad (5)$$

где d_{MIN} задает нижнюю границу числа чанков, в которых термин должен встретиться, что позволяет сохранять редкие, но предметно значимые лексемы, а d_{MAX} задает верхнюю границу относительной распространенности термина по корпусу, исключая из словаря чрезмерно частые, малоинформативные единицы.

Итоговый словарь V определяется как множество терминов, удовлетворяющих указанным условиям:

$$\mathrm{V} = \left\{ t \in \mathrm{V}' : df(t) \geq d_{\mathrm{MIN}}, \frac{df(t)}{N} \leq d_{\mathrm{MAX}} \right\}, \quad (6)$$

где V′ – совокупность всех униграмм и биграмм, извлеченных из корпуса чанков.

Таким образом, в семантико–онтологической модели признаков паспорта специальностей задают базисное пространство *n*-грамм V, в котором далее строятся TF-IDF-представления как самих паспортов (на уровне чанков и центроидов), так и входных документов.

Использование чанкования и комбинированного словаря униграмм и биграмм позволяет совместить локальную тематическую чувствительность с устойчивостью к шуму и обеспечить интерпретируемость получаемых признаковых векторов.

3.2. Лексический слой

На лексическом уровне каждый чанк паспорта специальности рассматривается как документ в общем словаре *n*-грамм V, построенном по корпусу паспортов (см. п. 3.1). Для каждого термина $t \in \mathrm{V}$ и чанка c вычисляется частота вхождений $\mathrm{tf}(t, c)$, после чего формируется взвешенное представление с использованием алгоритма TF-IDF.

Для учета повторных вхождений терминов в документ используется сублогарифмическая частота термина (сублогарифмический TF):

$$\mathrm{tf}(t,c) = \begin{cases} 1 + \ln(\mathrm{tf}(t,c)), & \text{при } \mathrm{tf}(t,c) > 0, \\ 0, & \text{при } \mathrm{tf}(t,c) > 0. \end{cases} \quad (7)$$

Использование сублогарифмического TF позволяет уменьшить вклад часто повторяющихся терминов внутри одного фрагмента, поскольку каждое последующее появление увеличивает их вес с уменьшающимся приростом, что снижает влияние локальных аномально высоких частот и уменьшает зависимость оценки от длины фрагмента текста.

Обратная частота IDF вычисляется по сглаженной формуле:

$$\mathrm{idf}(t) = \log \frac{1+N}{1+df(t)} + 1, \quad (8)$$

где N – общее число чанков в корпусе паспортов; $df(t)$ – количество чанков, в которых термин t встречается хотя бы один раз.

Добавление единицы в числитель и знаменатель обеспечивает численную устойчивость при крайних значениях $df(t)$, а сдвиг на единицу сохраняет положительность весов даже для терминов, встречающихся во всех документах. При этом функция $\mathrm{idf}(t)$ остается монотонно убывающей по мере роста $df(t)$, редкие термины получают более высокий вес, а часто встречающиеся термины – более низкий.

Вес термина t в чанке c задается произведением:

$$\mathrm{w}_t(c) = \mathrm{tf}(t, c) \cdot \mathrm{idf}(t). \quad (9)$$

Вектор $\mathbf{w}(c)$, значения которого заданы величинами $\mathrm{w}_t(c)$, интерпретируется как онтологический вектор, где каждая координата соответствует лексеме или устойчивому словосочетанию, ассоциированному с элементом онтологии (понятием, свойством или типичным отношением в предметной области).

Таким образом лексический слой фиксирует вклад онтологических маркеров, проявляющихся на уровне терминов и *n*-грамм.

Для обеспечения сопоставимости векторов выполняется L_2-нормализация:

$$\mathbf{x}(c) = \frac{\mathbf{w}(c)}{\| \mathbf{w}(c) \|_2}, \quad (10)$$

где $\|\mathbf{w}(c)\|_2$ обозначает длину вектора.

После L_2-нормализации учет общей длины и объема текста при сравнении документов существенно уменьшается, а направление вектора определяется относительным распределением весов терминов.

Поскольку значения TF-IDF неотрицательны, то скалярное произведение двух L_2-нормированных векторов совпадает с косинусным сходством и лежит в диапазоне [0; 1], что удобно для последующей интерпретации результатов.

Нормированные TF-IDF-векторы всех чанков паспортов формируют разреженную матрицу

$$\mathbf{X} \in \mathbb{R}^{M \times |\mathrm{V}|}, \quad (11)$$

где $M = \Sigma_s n_s$ – общее количество чанков по всем специальностям, $|\mathrm{V}|$ – размер словаря.

Матрица хранится в CSR[24]-формате, где отдельно сохраняются массив ненулевых весов, индексы соответствующих терминов и указатели границ строк. Использование CSR-формата обеспечивает компактное хранение и позволяет вычислять произведение $\mathbf{v}^{\mathrm{T}}\mathbf{X}$ за время, пропорциональное числу ненулевых компонент вектора v и ненулевых элементов матрицы $\mathbf{X}$.

В дальнейшем в том же словаре V и с теми же значениями $\mathrm{idf}(t)$ представляются входные документы (научные статьи, авторефераты диссертации). Их нормированные TF-IDF-векторы сопоставляются с векторами чанков и агрегированными представлениями паспортов, а косинусные сходства выступают базовыми информационно-онтологическими метриками в предлагаемой модели.

3.3. Семантический слой

Лексический слой фиксирует совпадения терминов и устойчивых словосочетаний, однако остается

[24] Compressed sparse row – формат хранения разреженных матриц.

чувствительным к вариативности формулировок, синонимии и аббревиатурам. Для учета контекстных связей между леммами вводится семантический слой, основанный на графовой модели ICAN, в которой каждый текст описывается в виде локального графа совместной встречаемости терминов [19].

Пусть нормализованный текст документа (чанка паспорта, статьи или автореферата) задан последовательностью лемм $d = (t_1, \ldots, t_T)$, а множество уникальных лемм этого текста обозначено как $W = \{u_1, ..., u_{m_d}\}$. Тогда для текста строится ориентированный взвешенный граф G_d на вершинах W, а его структура задается матрицей смежности $\mathbf{M} \in [0;1]^{m_d \times m_d}$, инициализируемой нулями.

Граф формируется при проходе по тексту со скользящим окном фиксированной ширины W (по умолчанию W = 11 токенов). В каждой позиции окна выбирается центральный токен x; все остальные токены в окне рассматриваются как его контекстные соседи y. Обновление матрицы **M** осуществляется в три этапа.

На первом этапе усиливаются прямые связи между центральным токеном и его контекстом (контекстными соседями слева и справа от центрального токена x).

Если связь $M_{xy} = 0$, то связь инициализируется базовым весом 0.5, а при повторных появлениях в том же контекстном отношении вес плавно увеличивается по формуле (12), что обеспечивает монотонный рост и при этом ограничивает величину веса.

$$M_{xy} = M_{xy} + \frac{1}{2}\left(1 - M_{xy}\right). \quad (12)$$

На втором этапе производится учет косвенных связей 2-го порядка. Если токен y уже связан с токеном k (т.е. $M_{yk} > 0$), то для пары (x, k) добавляется ослабленный вклад:

$$M_{yk} = M_{xk} + A(1 - M_{xk})M_{xy}M_{yk}, \quad (13)$$

где $A \ll 1$ – коэффициент масштабирования.

В результате полученный граф отражает не только совместную встречаемость терминов в одном скользящем окне, но и связи через общие контекстные соседи, что позволяет фиксировать более широкие контекстные ассоциации.

На третьем этапе к матрице **M** применяется операция затухания и пороговой фильтрации. Все веса умножаются на коэффициент $\gamma \in (0, 1)$, а элементы с величиной ниже порога θ принимаются равными нулю:

$$M_{xy} = \gamma M_{xy}, \quad (14)$$

где $M_{xy} = 0$ при $M_{xy} \le \theta$, $\gamma = 0.9$ – коэффициент затухания, $\theta = 0.4$ – пороговый коэффициент удаления связи.

Такая процедура подавляет случайные слабые связи и формирует более устойчивую структуру графа, отражающую стабильные контекстные связи.

В итоге семантическое представление текста в пространстве ICAN задается вектором степеней вершин. Для каждой леммы $u_i \in W$ вычисляется ее степень (суммарный вес исходящих и входящих ребер):

$$k_i = \Sigma_j M_{ij} + \Sigma_j M_{ji}. \quad (15)$$

Вектор $\mathbf{k}(d) = (k_1, ..., k_{m_d})$ отражает относительную важность лемм в контекстной структуре текста, где высокие значения соответствуют терминам, играющим роль «узловых точек» семантического графа.

Для интеграции с лексическим слоем и онтологическим основанием вектор $\mathbf{k}(d)$ проецируется в общий базис словаря V, построенного по корпусу паспортов. Если $V = \{v_1, \ldots, v_{|V|}\}$, то семантический вектор $\mathbf{s}(d) \in \mathbb{R}^{|V|}$ определяется по формуле:

$$s_j(d) = \begin{cases} k_i, & \text{если лемма } u_i \text{ совпадает с } v_j, \\ 0. \end{cases} \quad (16)$$

Далее выполняется L_2-нормализация:

$$\hat{\mathbf{s}}(d) = \frac{\mathbf{s}(d)}{\|\mathbf{s}(d)\|_2}, \quad (17)$$

что делает семантические векторы сопоставимыми по масштабу и позволяет использовать косинусное сходство.

4. МЕТРИКИ ИНФОРМАЦИОННОГО МОРФИЗМА

Обозначенные ранее в п. 3.2 лексический (TF-IDF) и в п. 3.3 семантический (ICAN) слои задают два согласованных признаковых пространства, в каждом из которых документ может быть сопоставлен с паспортами научных специальностей. В обоих случаях используются две группы метрик:

- локальные, измеряющие близость документа к отдельным фрагментам паспорта;
- глобальные, характеризующие близость документа к центроиду паспорта, полученному усреднением по всем его фрагментам.

Эти метрики вводятся в унифицированной форме и далее применяются отдельно в каждом слое.

Пусть для входного документа d построен нормированный вектор признаков $\mathbf{v}(d)$, а для каждого чанка $c_{s,i}$ паспорта s построен нормированный вектор $\mathbf{u}_{s,i}$, где $i = 1, \ldots, n_s$. Все векторы нормированы и неотрицательны.

Тогда локальная метрика задается как косинусное сходство документа с отдельным чанк-вектором паспорта по формуле:

$$\text{CosSimilarity}(d, c_{s,i}) = \langle \mathbf{v}(d), \mathbf{u}_{s,i} \rangle \in [0; 1]. \quad (18)$$

Для каждого паспорта s это определяет набор локальных оценок $\text{CosSimilarity}(d, c_{s,i})_{i=1}^{n_s}$, отражающих то, насколько сильно документ сопоставим по содержанию с отдельными фрагментами текста паспорта.

На основе этих оценок вводится агрегированная локальная метрика MaxSim_k, учитывающая только несколько наибольших совпадений

$$r_{s,i} = \text{CosSimilarity}(d, c_{s,i}),\ i = 1, \ldots, n_s, \quad (19)$$

а через $r_{s,1} \geq \ldots \geq r_{s,n_s}$ – упорядоченные по убыванию значения. Тогда локальная оценка для s задается следующим образом:

$$\text{MaxSim}_k(d, s) = \frac{1}{k_s} \sum_{j=1}^{k_s} r_{s,j},\ k_s = \min(k, n_s). \quad (20)$$

Таким образом, MaxSim_k отражает наличие в паспорте нескольких фрагментов, максимально близких к содержанию документа. Выбор параметра k обеспечивает компромисс между чувствительностью к узким совпадениям и устойчивостью к шуму и в экспериментальной части используется равный трем ($k = 3$).

Глобальная метрика формируется через центроид паспорта в рассматриваемом слое. Для паспорта s усредненный вектор определяется по формуле:

$$\bar{\mathbf{u}}_s = \frac{1}{n_s} \sum_{i=1}^{n_s} u_{s,i}. \quad (21)$$

После этого выполняется нормализация, а глобальная мера близости документа d к паспорту s задается косинусным сходством с центроидом по формуле:

$$\text{Centroid Cos}(d, s) = \left\langle \mathbf{v}(d), u_s \right\rangle = \frac{\sum_{i=1}^{n_s} \left\langle \mathbf{v}(d), u_s \right\rangle}{\left\| \sum_{i=1}^{n_s} u_s \right\|_2} \in [0;1]. \quad (22)$$

В результате в каждом слое документы и паспорта специальностей связаны парой комплементарных метрик: $\text{MaxSim}_k(d, s)$ – локальная метрика, чувствительная к наиболее релевантным фрагментам паспорта; $\text{CentroidCos}(d, s)$ – глобальная метрика, отражающая соответствие общей тематике.

4.1. Гибридная модель и интегральная метрика

Локальные и глобальные метрики, определенные для лексического и семантического слоев, задают согласованные оценки близости документа d к паспорту s.

Далее эти метрики объединяются в одно скалярное значение внутри соответствующего слоя, после чего результаты лексического и семантического слоев агрегируются в итоговую гибридную метрику.

Внутри каждого слоя локальная и глобальная метрики объединяются по формуле:

$$S(d, s) = (1 - \alpha)\text{CentroidCos}(d, s) + \alpha\text{MaxSim}(d, s), \quad (23)$$

где коэффициент α задает баланс между $\text{CentroidCos}(d, s)$, отражающей близость к паспорту в целом, и $\text{MaxSim}(d, s)$, выделяющей наиболее близкие фрагменты паспорта.

В лексическом слое TF-IDF увеличение α усиливает вклад точных терминологических совпадений, поскольку при $\alpha > 0.5$ возрастает влияние локальных максимумов MaxSim, возникающих в тех паспортах, которые содержат фрагменты с высокой концентрацией совпадающих n-грамм. Значение $\alpha = 0.6$ выбрано эмпирически по результатам валидации как обеспечивающее наилучшее качество классификации и рекомендационного ранжирования.

Полученные значения $S(d, s)$ для TF-IDF и ICAN интерпретируются как две независимые, но согласованные оценки соответствия документа паспорту. Первая – в терминах точных лексических маркеров, вторая – в терминах контекстных ассоциаций. Для формирования итогового скалярного скора вводится гибридная метрика:

$$S(d, s) = (1 - \lambda)S_{TF}(d, s) + \lambda S_{ICAN}(d, s), \lambda \in [0; 1], \quad (24)$$

где параметр λ регулирует вклад семантического слоя относительно лексического.

При λ, стремящейся к единице, доминируют графовые представления ICAN, что повышает устойчивость к перефразированиям и аббревиатурам. В проведенных экспериментах эмпирически подобраны значения $\alpha_{TF} \approx 0.6$, $\alpha_{ICAN} \approx 0.5$, $\lambda \approx 0.5$, обеспечивающие баланс между точностью и устойчивостью на корпусах авторефератов и статей.

С вычислительной точки зрения все составляющие метрики $S(d, s)$ получаются из двух векторно-матричных операций вида $\mathbf{r} = \mathbf{X} \times \mathbf{v}(d)$, где $\mathbf{X}$ – матрица чанк-векторов, а $\mathbf{r}$ – вектор косинусных сходств со всеми чанками всех паспортов.

Формально, с концептуальной точки зрения значение $S(d, s)$ можно трактовать как интенсивность информационного морфизма от документа d к онтологической сущности s. Так как чем выше $S(d, s)$,

тем более согласовано содержание текста воспроизводит лексико-семантический профиль соответствующей специальности.

4.2. Оценка вероятностного морфизм

Интегральная метрика $S(d, s)$ задает для каждого документа d и паспорта специальности $s \in O$ скалярную оценку степени соответствия. Для перехода от набора таких оценок к формальному информационному морфизму требуется построить вероятностное распределение по онтологическому множеству паспортов O.

Пусть для фиксированного документа d вычислены значения $S(d, s)_{s \in O}$, тогда на их основе вводится распределение вероятностей по схеме softmax[25] с температурой $\tau > 0$:

$$P(s \mid d, \tau) = \frac{e^{(S(d,s)-M)/\tau}}{\sum_{q \in O} e^{(S(d,s)-M)/\tau}}, \quad (25)$$

где $M = \max_{q \in O} S(d, q)$ используется для численной стабилизации (log-sum-exp нормировка). Вычитание M не влияет на относительные вероятности, но предотвращает переполнение при экспоненциальном преобразовании.

Параметр τ определяет степень концентрации распределения вероятностей:

- при $\tau \ll 1$ распределение становится более концентрированным, поскольку даже небольшие различия в значениях $S(d, s)$ приводят к выраженному доминированию одного или нескольких паспортов;
- при $\tau \approx 1$ применяется стандартное температурное масштабирование, при котором softmax сохраняет типичное поведение нормировки оценок;
- при $\tau > 1$ распределение становится более сглаженным, что удобно при анализе междисциплинарных и пограничных текстов, для которых характерно наличие нескольких сопоставимых по величине оценок.

Полученное распределение $P(s \mid d, \tau)$ задает отображение:

$$\mu: D \to \Delta(O), \ \mu(d) = P(s \mid d, \tau), \quad (26)$$

где D – множество документов, а $\Delta(O)$ – симплекс вероятностных мер на множестве паспортов специальностей.

Именно μ интерпретируется как информационный морфизм от текстового пространства к онтологическому пространству O, где по каждому документу сопоставляется распределение его тематического вклада по нормативно заданным областям.

Точечное решение задачи классификации при этом соответствует паспорту с максимальной вероятностью:

$$\hat{s}_1(d) = \arg\max_{s \in O} P(s \mid d, \tau). \quad (27)$$

5. ЭКСПЕРИМЕНТАЛЬНАЯ ОЦЕНКА

5.1. Результаты валидации модели на авторефератах соискателей

Валидация модели проводилась на корпусе из 124 авторефератов соискателей. Корпус охватывает шестнадцать диссертационных советов пяти организаций. Это МИРЭА – Российский технологический университет (РТУ МИРЭА)[26], Рязанский государственный радиотехнический университет им. В.Ф. Уткина (РГРТУ)[27], Федеральный исследовательский центр «Информатика и управление» Российской академии наук (ФИЦ ИУ РАН)[28], Институт проблем управления им. В.А. Трапезникова Российской академии наук (ФИЦ ИПУ РАН)[29] и Российский университет дружбы народов имени Патриса Лумумбы (РУДН)[30].

Для каждого совета учитывались авторефераты по шифрам, входящим в его область компетенции, например, 2.3.2, 2.3.5, 2.3.8 для совета Д24.2.326.09. В качестве эталонной метки для каждого документа принимался официальный шифр специальности, указанный в данных диссертационного совета.

Для трех конфигураций модели: лексической TF-IDF, семантической ICAN и гибридной оценивались две метрики.

Первая метрика – точность top 1, т.е. доля авторефератов, для которых паспорт с максимальным значением $S(d, s)$ совпадает с эталонным шифром.

Вторая метрика – точность top 3, т.е. доля авторефератов, для которых эталонный шифр входит в тройку паспортов с наибольшими значениями $S(d, s)$.

[25] Температурный softmax – функция нормализации, переводящая логиты z_i в распределение вероятностей, где параметр температуры $T > 0$ регулирует «резкость» этого распределения.

[26] https://www.mirea.ru/. Дата обращения 04.04.2026. / Accessed April 04, 2026. (In Russ.).

[27] https://rsreu.ru/. Дата обращения 04.04.2026. / Accessed April 04, 2026. (In Russ.).

[28] https://www.frccsc.ru/. Дата обращения 04.04.2026. / Accessed April 04, 2026. (In Russ.).

[29] https://www.ipu.ru/. Дата обращения 04.04.2026. / Accessed April 04, 2026. (In Russ.).

[30] https://www.rudn.ru/. Дата обращения 04.04.2026. / Accessed April 04, 2026. (In Russ.).

Агрегированные по всему корпусу значения демонстрируют устойчивое преимущество гибридной модели. Средняя по всем советам и документам точность top 1 для TF-IDF составляет около 0.58, для ICAN – около 0.57, для гибридной конфигурации – около 0.69. Для точности top 3 картина еще более выраженная. Модель TF-IDF дает в среднем около 0.74, ICAN – около 0.81, гибридная модель достигает порядка 0.90, т.е. в девяти случаях из десяти верный паспорт входит в тройку наиболее близких по значению $S(d, s)$.

Рассмотрение результатов по отдельным организациям показывает схожую картину. В выборках РТУ МИРЭА (табл. 1), где анализировались советы Д24.2.326.09, Д24.2.326.03, Д24.2.326.08 и Д24.2.326.10, средняя точность top 1 по авторефератам равна примерно 0.54 для TF-IDF и около 0.63 для ICAN. Гибридная модель повышает этот показатель до 0.75.

По метрике top 3 гибридная конфигурация достигает около 0.96, т.е. почти во всех случаях верный шифр присутствует среди трех наиболее вероятных.

Таблица 1. Результаты модели для авторефератов в РТУ МИРЭА

Совет Д24.2.326.09			Совет Д24.2.326.03		
Научные специальности 2.3.2, 2.3.5, 2.3.8			Научные специальности 1.4.7, 1.4.10		
Метод	Точность по top 1	Точность по top 3	Метод	Точность по top 1	Точность по top 3
TF-IDF	0.6	0.7	TF-IDF	0.5	0.63
ICAN	0.7	0.7	ICAN	0.5	0.75
Гибридный	0.8	1	Гибридный	0.63	0.88
Совет Д24.2.326.08			Совет Д24.2.326.10		
Научные специальности 1.2.2, 2.3.1			Научная специальность 5.2.3		
Метод	Точность по top 1	Точность по top 3	Метод	Точность по top 1	Точность по top 3
TF-IDF	0.33	0.67	TF-IDF	0.67	0.67
ICAN	0.67	0.67	ICAN	0.67	1
Гибридный	0.67	1	Гибридный	1	1

В РГРТУ (табл. 2), для советов Д24.2.375.01, Д24.2.375.02, Д24.2.375.03 и Д99.2.113.02, лексическая модель демонстрирует среднюю точность top 1 порядка 0.68 при значении около 0.61 для ICAN.

При этом семантический слой дает более высокую точность top 3, около 0.89 против 0.78 у TF-IDF.

Гибридная модель сочетает преимущества обеих конфигураций и выходит на среднюю точность top 1 около 0.75 и top 3 около 0.93.

Таблица 2. Результаты модели для авторефератов в РГРТУ

Совет Д24.2.375.01			Совет Д24.2.375.02		
Научные специальности 2.3.1, 2.3.5			Научные специальности 1.3.2, 1.3.11, 2.2.1		
Метод	Точность по top 1	Точность по top 3	Метод	Точность по top 1	Точность по top 3
TF-IDF	0.67	0.67	TF-IDF	0.67	0.83
ICAN	0.67	1	ICAN	0.5	0.83
Гибридный	0.67	0.67	Гибридный	0.5	0.83
Совет Д24.2.375.03			Совет Д99.2.113.02		
Научные специальности 2.2.11, 2.2.12, 2.2.13			Научная специальность 2.3.8		
Метод	Точность по top 1	Точность по top 3	Метод	Точность по top 1	Точность по top 3
TF-IDF	0.69	0.81	TF-IDF	0.67	0.67
ICAN	0.63	0.94	ICAN	0.67	0.67
Гибридный	0.81	1	Гибридный	1	1

В ФИЦ ИУ РАН (табл. 3) для советов Д24.1.224.04, Д24.1.224.03 и Д24.1.224.02 средняя точность top 1 составляет около 0.60 для TF-IDF и около 0.50 для ICAN. Гибридная модель повышает этот показатель до 0.70.

При этом по метрике top 3 TF-IDF и ICAN дают близкие значения порядка 0.83, а гибридная конфигурация выходит на уровень около 0.90.

Таблица 3. Результаты модели для авторефератов в ФИЦ ИУ РАН

Совет Д24.1.224.04			Совет Д24.1.224.03		
Научные специальности 2.3.2, 2.3.5, 2.3.6			Научные специальности 1.2.1, 1.2.3, 2.3.8		
Метод	Точность по top 1	Точность по top 3	Метод	Точность по top 1	Точность по top 3
TF-IDF	0.67	1	TF-IDF	0.56	0.78
ICAN	0.67	1	ICAN	0.39	0.83
Гибридный	1	1	Гибридный	0.61	0.89

Совет Д24.1.224.02		
Научные специальности 1.1.2, 1.1.6, 1.1.9		
Метод	Точность по top 1	Точность по top 3
TF-IDF	0.67	0.89
ICAN	0.67	0.78
Гибридный	0.78	0.89

В ФИЦ ИПУ РАН (табл. 4), где корпус включает пять авторефератов совета Д24.1.107.02, выборка мала и результаты следует интерпретировать с осторожностью. В этой группе семантическая модель ICAN дает лучшую точность top 1 около 0.60 по сравнению с 0.40 у TF-IDF.

Гибридная конфигурация сохраняет уровень top 1 около 0.60, однако по top 3 не превосходит семантическую ветвь.

Таблица 4. Результаты модели для авторефератов в ФИЦ ИПУ РАН

Совет Д24.1.107.02		
Научные специальности 2.3.1, 2.3.4		
Метод	Точность по top 1	Точность по top 3
TF-IDF	0.4	0.6
ICAN	0.6	0.8
Гибридный	0.6	0.6

В РУДН (табл. 5), для четырех советов ПДС 0300.004, ПДС 2028.001, ПДС 0900.006 и ПДС 0200.002, средняя по корпусу точность top 1 для TF-IDF и ICAN близка и составляет около 0.54.

Гибридная модель повышает этот показатель до 0.62. Для top 3 лексическая и семантическая конфигурации дают значения порядка 0.70 и 0.78, в то время как гибридная модель достигает около 0.89. На отдельных советах наблюдаются существенные выигрыши.

Для ПДС 2028.001 гибридная модель улучшает точность top 3 с 0.56 у TF-IDF и 0.78 у ICAN до единицы. Для ПДС 0200.002 гибридная конфигурация также поднимает точность top 3 до единицы при сохранении высокой точности top 1.

Таким образом валидация на корпусе авторефератов соискателей показывает, что предложенная семантико-онтологическая модель в гибридной конфигурации обеспечивает устойчивое и интерпретируемое качество восстановления шифров специальностей.

В большинстве случаев верный паспорт попадает в узкий набор наиболее вероятных кандидатов, что делает модель пригодной как для автоматизированной поддержки выбора шифра, так и для анализа альтернативных тематически близких областей.

Таблица 5. Результаты модели для авторефератов в РУДН

Совет ПДС 0300.004			Совет ПДС 2028.001		
Научные специальности 3.1.18, 3.1.20, 3.3.6			Научные специальности 5.8.1, 5.8.7		
Метод	Точность по top 1	Точность по top 3	Метод	Точность по top 1	Точность по top 3
TF-IDF	0.56	0.75	TF-IDF	0.44	0.56
ICAN	0.44	0.81	ICAN	0.44	0.78
Гибридный	0.56	0.81	Гибридный	0.56	1
Совет ПДС 0900.006			Совет ПДС 0200.002		
Научная специальность 5.1.4			Научные специальности 1.4.1, 1.4.3, 1.4.4		
Метод	Точность по top 1	Точность по top 3	Метод	Точность по top 1	Точность по top 3
TF-IDF	0.4	0.6	TF-IDF	0.71	0.86
ICAN	0.8	0.8	ICAN	0.71	0.71
Гибридный	0.8	0.8	Гибридный	0.71	1

5.2. Результаты валидации модели на научных работах, опубликованных в Перечне ВАК

Валидация модели на научных статьях проводилась на корпусе публикаций из журналов, включенных в действующий Перечень ВАК. Для этих журналов известна их отнесенность к укрупненным областям наук и к группам научных специальностей, что позволило провести оценку как на агрегированном уровне областей, так и на уровне групп паспортов.

В первом эксперименте анализировалась точность отнесения статьи к укрупненной области наук. Эталонная область определялась по профилю журнала в Перечне ВАК, предсказанная область по паспорту с максимальным значением $S(d, s)$. Результаты показывают различия в поведении лексической, семантической и гибридной конфигураций (табл. 6).

Таблица 6. Результаты оценки научных статей по укрупненным областям наук

Метод	Точность по каждой укрупненной области наук				
	1. Естественные науки	2. Технические науки	3. Медицинские науки	4. Сельскохозяйственные науки	5. Социальные и гуманитарные науки
TF-IDF	0.85	0.87	0.85	0.86	0.83
ICAN	0.79	0.72	0.74	0.9	0.91
Гибридный	0.9	0.88	0.93	0.92	0.94

Метод TF-IDF демонстрирует наибольшую устойчивость в естественных и технических науках, где точность top 1 составляет соответственно 0.85 и 0.87. В медицинских и сельскохозяйственных науках точность TF-IDF равна 0.85 и 0.86, в социально-гуманитарных науках – 0.83. Это согласуется с тем, что в естественнонаучных и технических областях терминология более стандартизована и ближе к формулировкам паспортов.

Семантическая модель ICAN заметно выигрывает там, где язык публикаций более вариативен. В социально-гуманитарных науках точность достигает 0.91 при 0.83 у TF-IDF. В сельскохозяйственных науках ICAN дает 0.90 против 0.86 у TF-IDF. В естественных, технических и медицинских науках ICAN немного уступает лексической модели, что отражает зависимость чисто семантического слоя от качества локальных графов в терминонасыщенных, но хорошо стандартизованных областях.

Гибридная конфигурация объединяет сильные стороны обоих подходов. По всем пяти укрупненным областям она дает наибольшие значения точности top 1. Для естественных наук точность составляет 0.90, для технических – 0.88, для медицинских – 0.93, для сельскохозяйственных – 0.92, для социально-гуманитарных – 0.94. Таким образом, при переходе к более высокому уровню агрегирования гибридная модель практически всегда исправляет ошибки каждой из одиночных конфигураций и обеспечивает наиболее устойчивое поведение.

Во втором эксперименте оценивалась точность восстановления групп научных специальностей. Для каждой статьи формировался вектор S(d, s) по всем паспортам, после чего предсказанная группа определялась по паспорту с максимальным значением S(d, s). Эталонная группа задавалась по заявленной специализации журнала. Здесь анализировались метрики top 1 и top 3 (табл. 7).

Таблица 7. Результаты оценки научных статей по группе научных специальностей

Метод	Точность по top 1	Точность по top 3
TF-IDF	0.9	0.94
ICAN	0.87	0.97
Гибридный	0.91	0.96

На уровне групп специальностей лексическая модель TF-IDF достигает точности top 1, равной 0.90, и точности top 3, равной 0.94. Семантическая модель ICAN дает немного меньшую точность top 1, равную 0.87, но более высокую точность top 3, равную 0.97. Это означает, что ICAN несколько чаще ошибается в выборе единственного наиболее близкого паспорта, но почти всегда включает правильную группу в число трех наиболее вероятных.

Гибридная модель сохраняет наилучшее значение top 1, равное 0.91, и при этом дает высокую точность top 3, равную 0.96. Она не превосходит ICAN по top 3, однако обеспечивает более сбалансированное соотношение между точностью первого выбора и полнотой тройки ближайших групп. В контексте задач рекомендационного ранжирования по шифрам такая конфигурация является наиболее практичной, т.к. позволяет одновременно надежно предлагать основной код и формировать содержательно релевантный список альтернативных специальностей.

В совокупности результаты на статьях из журналов Перечня ВАК подтверждают выводы, сделанные по авторефератам. Лексическая модель лучше работает в областях с жестко закрепленной номенклатурой, семантическая модель особенно полезна в гуманитарных и близких к ним доменах, гибридная конфигурация дает наиболее устойчивое качество на всех уровнях агрегирования и обеспечивает интерпретируемые вероятностные профили по группам научных специальностей.

5.3. Сравнение графовой семантико-онтологической модели с большими языковыми моделями

Для оценки предложенной модели проведено сравнение с рядом крупных языковых моделей, примененных в режиме классификации без дообучения. Во всех случаях ставилась одна и та же задача восстановить шифр специальности или ближайший паспорт по тексту документа. На корпусах авторефератов и научных статей рассчитывались точность top 1 и точность top 3 (табл. 8).

По авторефератам гибридная графовая семантико-онтологическая модель дает точность top 1, равную 0.69, и точность top 3, равную 0.90. Лексическая TF-IDF-модель и семантическая ICAN уступают ей по обеим метрикам (для TF-IDF – 0.58 и 0.74, для ICAN – 0.57 и 0.81). Среди больших языковых моделей наилучшие значения показывает конфигурация ChatGPT 5.2 Thinking, у которой точность top 1 достигает 0.79, а точность top 3 – 0.84. Модель ChatGPT 4o[31] работает на уровне 0.71 и 0.73, DeepSeek[32] – на уровне 0.61 и 0.70, LLaMA[33] – на уровне 0.57 и 0.63, Алиса AI (YandexGPT)[34] заметно

[31] https://openai.com/ru-RU/index/hello-gpt-4o/. Дата обращения 04.04.2026. / Accessed April 04, 2026. (In Russ.).

[32] https://www.deepseek.com/. Дата обращения 04.04.2026. / Accessed April 04, 2026.

[33] https://www.llama.com/. Дата обращения 04.04.2026. / Accessed April 04, 2026.

[34] https://alice.yandex.ru/. Дата обращения 04.04.2026. / Accessed April 04, 2026. (In Russ.).

уступает всем вариантам и дает 0.46 и 0.52. Таким образом, по авторефератам крупная языковая модель ChatGPT 5.2 Thinking[35] превосходит гибридную схему по точности первого выбора, однако графовая семантико-онтологическая модель обеспечивает более высокую полноту по тройке ближайших специальностей и формирует размытый, но содержательно устойчивый профиль близостей по паспортам.

По научным статьям из журналов Перечня ВАК картина иная. Гибридная модель достигает точности top 1, равной 0.91, и точности top 3, равной 0.93. Лексическая TF-IDF модель дает точность 0.85 и 0.86, ICAN – 0.81 и 0.84. Среди больших языковых моделей ChatGPT 4o показывает точность 0.80 и 0.82, ChatGPT 5.2 Thinking – 0.82 и 0.97, DeepSeek – 0.79 и 0.84, LLaMA – 0.62 и 0.67, Алиса AI – 0.56 и 0.69. По метрике top 1 гибридная модель опережает все языковые модели с отрывом от девяти до пятнадцати процентных пунктов. По метрике top 3 лучший результат показывает ChatGPT 5.2 Thinking (0.97), гибридная модель дает немного меньшую величину, равную 0.93, но остается заметно выше остальных конфигураций. Это означает, что при классификации статей предложенная графовая семантико-онтологическая схема лучше фиксирует основной паспорт, в то время как крупные языковые модели чаще включают правильный шифр в широкий набор ближайших кандидатов.

Различия в поведении хорошо согласуются с природой сравниваемых подходов. Графовая семантико-онтологическая модель жестко привязана к текстам паспортов специальностей, использует их как онтологическое основание и факторизует решения на интерпретируемые компоненты TF-IDF и ICAN.

Это дает эффект, особенно заметный на статьях из журналов Перечня ВАК, где формулировки ближе к нормативному языку паспортов. Крупные языковые модели опираются на обобщенные представления о научных дисциплинах и часто учитывают контекст, который не отражен в текстах паспортов, поэтому на авторефератах, в которых присутствуют развернутые обзоры, ссылки на смежные области и менее формализованное изложение, лучшие конфигурации ChatGPT демонстрируют более высокую точность первого выбора, однако теряют часть интерпретируемости и управляемости.

Сравнение показывает, что предложенная графовая семантико-онтологическая модель сопоставима с современными языковыми моделями по качеству классификации и превосходит их по ряду метрик в сценариях, где важна строгая согласованность с текстами паспортов, при этом остается существенно дешевле по вычислительным затратам и прозрачнее по структуре принимаемых решений.

Крупные языковые модели целесообразно рассматривать как дополнительный инструмент, который дополняет, а не заменяет онтологически ориентированную схему, особенно в задачах экспертной поддержки, где требуется одновременно количественный скор и явная привязка результата к нормативным описаниям научных специальностей.

6. ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ И ЗАКЛЮЧЕНИЕ

Представленная в работе графовая семантико-онтологическая модель анализа научных текстов основана на паспортах научных специальностей ВАК, которые используются как текстовые описания предметных областей и одновременно как источник онтологического признакового пространства. На основе текстов паспортов формируется единый словарь

Таблица 8. Результаты сравнения графовой семантико-онтологической модели с большими языковыми моделями

Авторефераты			Статьи		
Метод	Точность по top 1	Точность по top 3	Метод	Точность по top 1	Точность по top 3
TF-IDF	0.58	0.74	TF-IDF	0.85	0.86
ICAN	0.57	0.81	ICAN	0.81	0.84
Гибридный	0.69	0.9	Гибридный	0.91	0.93
ChatGPT 4o	0.71	0.73	ChatGPT 4o	0.8	0.82
ChatGPT 5.2 Thinking	0.79	0.84	ChatGPT 5.2 Thinking	0.82	0.97
Алиса AI (YandexGPT)	0.46	0.52	Алиса AI (YandexGPT)	0.56	0.69
LLaMA	0.57	0.63	LLaMA	0.62	0.67
DeepSeek	0.61	0.7	DeepSeek	0.79	0.84

[35] https://openai.com/ru-RU/index/introducing-gpt-5-2/. Дата обращения 04.04.2026. / Accessed April 04, 2026. (In Russ.).

лемм и устойчивых словосочетаний, затем в этом базисе описываются сами паспорта, авторефераты диссертаций и научные статьи.

Значение сходства документа с шифром научной специальности интерпретируется через согласованность содержания текста с нормативным описанием области.

В то же время семантико-онтологическое пространство имеет два согласованных слоя. Лексический слой строится на основе векторов TF-IDF, которые фиксируют использование терминов и терминологических сочетаний, отобранных по корпусу чанков паспортов. Семантический слой основан на модели ICAN и описывает текст через граф совместной встречаемости лемм и степени его вершин. Это позволяет учитывать перефразирование, синонимию, использование аббревиатур и термины, разнесенные по тексту, при сохранении привязки к одному и тому же онтологическому базису.

Гибридная метрика, основанная на лексическом и семантическом подходах, объединяет локальные и глобальные метрики в каждом из слоев.

Экспериментальные результаты на корпусах авторефератов показывают, что гибридная модель по точности систематически превосходит отдельные лексические и семантические подходы. Лексическая модель на основе TF-IDF дает высокую точность там, где терминология стандартизована и хорошо совпадает с формулировками паспортов, а семантическая модель особенно полезна в областях с более свободным научным стилем, активным использованием сокращений и вариативных описаний предмета исследования.

Агрегированные показатели по укрупненным областям знаний подтверждают наблюдаемую картину. Для социально-гуманитарных направлений вклад семантического слоя оказывается критичным для повышения качества ранжирования, поскольку тематические границы часто выражаются через сложные формулировки и контекстуальные маркеры. Для значительной части технических и естественно-научных специальностей лексический слой уже обеспечивает высокий базовый уровень качества, а семантический слой уточняет профиль документа в случае близких шифров и пограничных тематик.

Таким образом модель демонстрирует содержательно объяснимое поведение, которое согласуется с особенностями терминологической структуры разных научных областей.

Дальнейшее развитие модели может быть с несколькими направлениями. Одно из них заключается в формировании лексем и терминологических сочетаний паспортов с внешними онтологиями, базами знаний и специализированными терминологическими ресурсами, что позволит перейти от текстового онтологического основания к более формализованной мультидоменной онтологии. Кроме того, возможно направление с добавлением контекстных эмбеддингов и нейросетевых моделей в виде дополнительного слоя над существующими TF-IDF и ICAN при сохранении интерпретируемости через разложение по онтологическому базису.

Отдельную задачу представляет оптимизация параметров на валидационных выборках и анализ их зависимости от области знания и типа документа.

В заключение можно отметить, что графовая семантико-онтологическая модель, основанная на текстах паспортов специальностей ВАК, демонстрирует возможность превращения нормативного корпуса в рабочее онтологическое пространство для количественного анализа научных текстов. Полученные результаты для авторефератов и научных статей показывают, что такой подход может служить основой как для автоматизированной поддержки принятия экспертных решений, так и для мониторинга структуры научного знания в рамках заданной нормативной онтологии.

Вклад авторов

Н.С. Курдюков – концепция исследования, методология, формальный анализ, написание первоначального варианта рукописи. Разработал основную концепцию информационного морфизма в рамках семантико-онтологических и графовых структур, разработал методологический подход, реализовал основные алгоритмы, провел формальные эксперименты и анализ, а также подготовил первоначальный проект рукописи.

В.Н. Калинин – методология, разработка программного обеспечения, обработка данных, проверка результатов, написание первоначального варианта рукописи. Внес вклад в совершенствование методологической базы, разработал вычислительные инструменты, провел валидацию разработанных моделей и инструментов на экспериментальных наборах данных, курировал и структурировал текстовые материалы, использованные в исследовании, и подготовил первоначальный проект рукописи.

С.А. Кудж – методологический надзор; проверка результатов; переработка и редактирование рукописи. Осуществлял экспертный надзор за методологической базой, обеспечивал научную обоснованность и строгость предложенных моделей и аналитических процедур, а также участвовал в критическом рассмотрении и доработке рукописи.

Д.О. Жуков – авторский надзор; методологический надзор; проверка достоверности; переработка и редактирование рукописи. Осуществлял общее научное руководство исследованием, следил за развитием и последовательностью методологического подхода, подтверждал результаты исследования на концептуальном и теоретическом уровнях, а также критически рассматривал и одобрял окончательный вариант рукописи.

Authors' contributions

N.S. Kurdyukov – conceptualization; methodology; formal analysis; writing the original draft. He has developed the core concept of information morphism within semantic-ontological and graph-based frameworks, designed the methodological approach, implemented the primary algorithms, conducted formal experiments and analysis, and prepared the initial manuscript draft.

V.N. Kalinin – methodology; validation; data curation; software; writing the original draft. He has contributed to the refinement of the methodological framework, developed computational tools, performed validation of the developed models and tools on experimental datasets, curated and structured the textual corpora used in the study, and prepared initial manuscript draft.

S.A. Kudzh – methodology supervision; validation; writing the review and editing. He has provided expert supervision of the methodological framework, ensured the scientific validity and rigor of the proposed models and analytical procedures, and contributed to critical review and refinement of the manuscript.

D.O. Zhukov – supervision; methodology oversight; validation; writing the review and editing. He has led the overall scientific supervision of the study, oversaw the development and consistency of the methodological approach, validated the research outcomes at a conceptual and theoretical level, and critically reviewed and approved the final manuscript.

СПИСОК ЛИТЕРАТУРЫ / REFERENCES

1. Altınel B., Ganiz M.C. Semantic text classification: A survey of past and recent advances. *Inf. Process. Management*. 2018;54(6):1129–1153. https://doi.org/10.1016/j.ipm.2018.08.001
2. Sikelis K., Tsekouras G.E., Kotis K.I. Ontology-based Feature Selection: A Survey. *arXiv* preprint arXiv:2104.07720 [cs.AI], 2021. https://doi.org/10.48550/arXiv.2104.07720
3. Ehring D., Ferraz-Doughty P., Luttmer J., Nagarajah A. A first step towards automatic identification and provision of user-specific knowledge: A verification of the feasibility of automatic text classification using the example of standards. *Procedia CIRP*. 2023;119:1103–1108. https://doi.org/10.1016/j.procir.2023.02.183
4. Layer M., Luttmer J., Nagarajah A., Stelzer R. Structured representation of pre-defined information backflow in standards and directives. *Standards*. 2024;4:262–285. https://doi.org/10.3390/standards4040013
5. Stănescu G., Oprea S.-V. Recent trends and insights in semantic web and ontology-driven knowledge representation across disciplines using topic modeling. *Electronics*. 2025;14(7):1313. https://doi.org/10.3390/electronics14071313
6. Touza I., Balama G., Lazarre W., Guidedi K., Kolyang. Ontology-driven text classification and data mining: Beyond keywords toward semantic intelligence. *Revue d'Intelligence Artificielle*. 2025;39(3):25–35. https://doi.org/10.18280/ria.390301
7. Pertsas V., Constantopoulos P. Ontology-driven extraction of contextualized information from research publications. In: *Proceedings of the 15th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management* (*IC3K 2023*). V. 2. KEOD. 2023. P. 108–118. https://doi.org/10.5220/0012254100003598
8. Mohd M., Javeed S., Nowsheena, Wani M.A., Khanday H.A. Sentiment analysis using lexico-semantic features. *J. Inform. Sci*. 2024;50(6):1449–1470. https://doi.org/10.1177/01655515221124016
9. Demidova L., Zhukov D., Andrianova E., Kalinin V. Model of lexico-semantic bonds between texts for creating their similarity metrics and developing statistical clustering algorithm. *Algorithms*. 2023;16:198. https://doi.org/10.3390/a16040198
10. Saeeda L., Med M., Ledvinka M., Blaško M., Křemen P. Entity linking and lexico-semantic patterns for ontology learning. In: Harth A., et al. *The Semantic Web*. Series: Lecture Notes in Computer Science. 2020. V. 12123. P. 138–153. https://doi.org/10.1007/978-3-030-49461-2_9
11. Yelmen I., Gunes A., Zontul M. Multi-class document classification using lexical ontology-based deep learning. *Appl. Sci*. 2023;13(10):6139. https://doi.org/10.3390/app13106139
12. Bugueño M., de Melo G. Connecting the dots: What graph-based text representations work best for text classification using graph neural networks? In: *Findings of the Association for Computational Linguistics*: *EMNLP 2023*. 2023. P. 8943–8960. https://doi.org/10.18653/v1/2023.findings-emnlp.600
13. Varella Ehrenfried H., Venturi Date V.T., Todt E. Exploring graph representation strategies for text classification. *Connect. Sci*. 2023;35(1):2289832. https://doi.org/10.1080/09540091.2023.2289832
14. Sánchez-Antonio C., Valdez-Rodríguez J.E., Calvo H. TTG-Text: A graph-based text representation framework enhanced by typical testors for improved classification. *Mathematics*. 2024;12:3576. https://doi.org/10.3390/math12223576
15. Onan A. Hierarchical graph-based text classification framework with contextual node embedding and BERT-based dynamic fusion. *Journal of King Saud University – Computer and Information Sciences.* 2023;35(7):101610. https://doi.org/10.1016/j.jksuci.2023.101610
16. Цветков В.Я., Курдюков Н.С. Информационное онтологическое моделирование. *Russian Technological Journal*. 2025;13(2):18–26. https://doi.org/10.32362/2500-316X-2025-13-2-18-26
[Tsvetkov V.Ya., Kurdyukov N.S. Informational ontological modeling. *Russian Technological Journal.* 2025;13(2):18–26. https://doi.org/10.32362/2500-316X-2025-13-2-18-26]
17. Nabhan A.R., Shaalan K. A graph-based approach to text genre analysis. *Computación y Sistemas*. 2016;20(3):527–539. https://doi.org/10.13053/CyS-20-3-2471

18. Ali I., Melton A. Semantic-based text document clustering using cognitive semantic learning and graph theory. In: *Proceedings of the 12th IEEE International Conference on Semantic Computing* (*ICSC 2018*). 2018. P. 243–247. https://doi.org/10.1109/ICSC.2018.00042
19. Lemaire B., Denhière G. Incremental construction of an associative network from a corpus. In: *Proceedings of the 26th Annual Meeting of the Cognitive Science Society*. 2004. V. 26. P. 825–830.

Об авторах

Курдюков Никита Сергеевич, аспирант, кафедра инструментального и прикладного программного обеспечения, Институт информационных технологий, ФГБОУ ВО «МИРЭА – Российский технологический университет» (119454, Россия, Москва, пр-т Вернадского, д. 78). E-mail: nskurdyukov@gmail.com. SPIN-код РИНЦ 8535-1612, https://orcid.org/0000-0001-6784-3369

Калинин Владимир Николаевич, ассистент, кафедра телекоммуникаций, Институт радиоэлектроники и информатики, ФГБОУ ВО «МИРЭА – Российский технологический университет» (119454, Россия, Москва, пр-т Вернадского, д. 78). E-mail: kalinin_v@mirea.ru. Scopus Author ID 57562579000, https://orcid.org/0000-0003-1365-4639

Кудж Станислав Алексеевич, д.т.н., профессор, профессор кафедры инструментального и прикладного программного обеспечения, Институт информационных технологий, ФГБОУ ВО «МИРЭА – Российский технологический университет» (119454, Россия, Москва, пр-т Вернадского, д. 78). E-mail: kudzh@mirea.ru. Scopus Author ID 56521711400, ResearcherID AAG-1319-2019, SPIN-код РИНЦ 8173-1572, https://orcid.org/0000-0003-1407-2788

Жуков Дмитрий Олегович, д.т.н., профессор, профессор кафедры телекоммуникаций, Институт радиоэлектроники и информатики, ФГБОУ ВО «МИРЭА – Российский технологический университет» (119454, Россия, Москва, пр-т Вернадского, д. 78). E-mail: zhukov_do@mirea.ru. Scopus Author ID 57189660218, SPIN-код РИНЦ 1798-8891, https://orcid.org/0000-0002-1211-5214

About the Authors

Nikita S. Kurdyukov, Postgraduate Student, Department of Instrumental and Applied Software, Institute of Information Technologies, MIREA – Russian Technological University (78, Vernadskogo pr., Moscow, 119454 Russia). E-mail: nskurdyukov@gmail.com. RSCI SPIN-code 8535-1612, https://orcid.org/0000-0001-6784-3369

Vladimir N. Kalinin, Assistant, Department of Telecommunications, Institute of Radio Electronics and Informatics, MIREA – Russian Technological University (78, Vernadskogo pr., Moscow, 119454 Russia). E-mail: kalinin_v@mirea.ru. Scopus Author ID: 57562579000, https://orcid.org/0000-0003-1365-4639

Stanislav A. Kudzh, Dr. Sci. (Eng.), Professor, Department of Instrumental and Applied Software, Institute of Information Technologies, MIREA – Russian Technological University (78, Vernadskogo pr., Moscow, 119454 Russia). E-mail: kudzh@mirea.ru. Scopus Author ID 56521711400, ResearcherID AAG-1319-2019, RSCI SPIN-code 8173-1572, https://orcid.org/0000-0003-1407-2788

Dmitry O. Zhukov, Dr. Sci. (Eng.), Professor, Department of Telecommunications, Institute of Radio Electronics and Informatics, MIREA – Russian Technological University (78, Vernadskogo pr., Moscow, 119454 Russia). E-mail: zhukov_do@mirea.ru. Scopus Author ID 57189660218, RSCI SPIN-code 1798-8891, https://orcid.org/0000-0002-1211-5214