Information systems. Computer sciences. Issues of information security

Информационные системы. Информатика. Проблемы информационной безопасности

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

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[1] 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.

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[2], 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

[1] 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] 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.).

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.

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[3]. 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[4] представления и локальные графы ICAN[5]. 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[6]. Качество модели оценивается на корпусе авторефератов и статей из журналов Перечня ВАК РФ[7], дополнительно проводится сравнение с крупными языковыми моделями.

[3] 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.).]

[4] 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.]

[5] 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.]

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

[7] 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.).]

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.

INTRODUCTION

While classical text classification methods based on the bag-of-words model and simple vector representations provide quite acceptable results under conditions of strict terminological standardization, such methods weakly take into account synonymy, variability of wording, and contextual relationships between terms. Modern neural network and graph-based text processing models partially eliminate these limitations; however, such models are typically trained on general corpora and are thus weakly tied to specific normative ontologies, which reduces the interpretability of the result. As will be discussed later in the article, this limitation presents particular challenges when addressing problems related to the processing of official documents based on standardization.

For example, systems for assessing scientific personnel and reviewing publications rely on regulations for the subject-specific classification of works into scientific specialties. In the Russian context, this role is fulfilled by the specialty passports of the Higher Attestation Commission (VAK in Russian transliteration),[8] which describe research objects, typical tasks, and methods to define the normative subject space. Due to the rapid growth in the volume of digital scientific texts, a need arises for automated methods used to compare articles and dissertations with these normative descriptions. Such methods are necessary for assigning specialty codes to dissertations and journal articles, analyzing the subject profile of dissertation councils and journals, and monitoring the overall structure of scientific activity.

One of the potential solutions to these problems consists in the development of precise, yet technologically advanced and rapid tools based on the use of ontological methods and approaches that have already become classical.

In this article, we implement an ontological approach to solving practical text analysis problems on the example of VAK specialty passports. Providing an ontological foundation for creating the necessary analysis tools, these passports form a unified feature space that describes both the passports themselves and scientific texts. The proposed graph semantic-ontological feature model created on their basis combines a term frequency—inverse document frequency (TF-IDF)[9] lexical layer with a semantic layer based on local context graphs of incremental construction of an associative network (ICAN).[10] For each document and passport, local and global proximity metrics are determined and combined into a hybrid scalar score. The translation of this score into a probabilistic profile based on codes is interpreted as an information morphism: a structure-preserving mapping that transfers the document representation from the feature space to the ontological space of passports in such a way that lexical and context-semantic relationships are consistently represented in terms of the normative concepts of the specialties.

[8] 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 (TF-IDF) is a statistical measure that takes into account the term frequency in a document and the inverse document frequency.

[10] Incremental construction of an associative network (ICAN) is a computational model for incremental construction of an associative network based on a text corpus.

This design approach solves two problems simultaneously. On the one hand, it maintains strict adherence to the standard language of the passports to ensure transparent interpretation of the contribution of individual terms and contextual relationships. On the other hand, the use of a graph semantic layer and hybrid aggregation allows variability in wording, abbreviations, and interdisciplinary connections to be taken into account. This enables not only targeted code selection, but also the analysis of the distribution of a document's thematic contribution across several related specialties, and the construction of aggregated profiles of scientific actors.

The purpose of the present work is to develop and experimentally evaluate a graph-based semantic-ontological model for analyzing Russian-language scientific texts based on the VAK specialty classifications. The model is then compared with large language models for the purposes of specialty code identification and subject-matter analysis. After compiling a corpus of classifications and accompanying texts, a unified normalization pipeline is developed, lexical and semantic feature layers are defined to produce an integrated information morphism metric, and validation is carried out using abstracts of applicants and articles from journals on the VAK List.[11]

The work is organized as follows. The first section provides an overview of works on semantic text classification, ontology-oriented models, and graph representations. The second section describes the ontological foundation of the study and the data corpus. The third section introduces a graph semantic-ontological feature model based on TF-IDF and ICAN. The fourth section formalizes information morphism metrics and a probabilistic interpretation of the results. The fifth section presents the results of validation on abstracts and scientific articles and a comparison with large language models. The sixth section discusses the obtained results, formulates conclusions, and formulates directions for further research.

1. LITERATURE REVIEW

The review paper [1] systematizes approaches to semantic text classification and compares them with traditional methods based on the bag-of-words (BoW) method. Demonstrating the limitations of BoW vector representation (high dimensionality, sparsity, ignoring synonymy and polysemy), the authors identify five main semantic methods: ontology-based, corpus-based, deep learning models, word/symbol sequence-based, and linguistically enriched approaches. The review summarizes the results of numerous experiments and demonstrates the advantage of semantic models in classification accuracy.

The paper [2] analyzes knowledge extraction methods in the context of the Semantic Web, with an emphasis on ontology-based feature selection. The use of ontologies to represent and select features in classification problems reduces dimensionality and increases the interpretability of models. However, the quality of the approach depends significantly on the completeness and consistency of the ontologies.

In [3–5], ontologies and semantic technologies are considered as a key tool for structuring engineering and scientific knowledge. In [3], an ontology-oriented approach to the automatic classification of engineering standards is proposed, using domain ontologies to map document fragments to business units.

In [4], a two-stage method for describing predefined information flows in standards and directives is developed, including a generalized data model and a machine-readable representation, which increases the accessibility and reusability of information in digital product models. The work [5] contains a bibliometric and semantic analysis of research on ontologies and the Semantic Web, highlighting the main areas related to dynamic updating of ontologies and scalability in the context of the Big Data.

The paper [6] presents a review of ontology-oriented methods for text classification, in which ontologies are considered as a formal basis for enriching vector representations and deep learning architectures through explicit modeling of domain concepts and their relationships. The study [7] proposes an ontology-oriented approach to extracting contextualized information from scientific publications. Based on transformer models, a two-stage pipeline is implemented, including sentence classification and entity recognition (research activities and methods). Relationships between activities and methods are then extracted via integration of the obtained data and publication metadata in the form of a resource description framework (RDF) graph.

A number of works [8–10] demonstrate the potential of lexical-semantic representations of text for solving various analysis problems. In [8], special lexical-semantic parameters for determining sentiment are introduced. These are obtained through the semantic expansion of sentiment lexicons and distributed representations, which unifies data dimensionality to improve classification quality.

The authors of the work [9] propose a model of lexical-semantic connections between documents and a statistical clustering algorithm based on cosine

[11] 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.).

proximity in the lexical-semantic space, comparable in quality to the Affinity Propagation method[12].

The work [10] uses a set of lexical-semantic patterns and procedures for matching textual mentions with domain ontology objects for their semi-automatic replenishment. Taken together, these approaches confirm the effectiveness of lexical-semantic features.

In [11], researchers address the problem of multi-class classification of imbalanced text data using the WordNet[13] lexical ontology and the bidirectional encoder representation from transformers (BERT)[14] model. The WordNet lexical ontology is used for ontology-based feature dimensionality reduction, followed by classification using traditional algorithms and a pre-trained BERT model. Experiments on a seven-class corpus show that a combination of WordNet and BERT provides the highest accuracy (up to 93.77%) as compared with variants without ontological feature refinement.

A number of works [12–15] systematically investigate graph representations of text and graph neural network models for classification problems. In [12], a large-scale comparison of various graph construction schemes and graph neural network (GNN) architectures with transformer models is conducted. In [13], researchers analyze how the choice of a graph text representation strategy affects the quality of GNN models. Works [14] and [15] develop the use of graph models in text content analysis architectures, focusing on improving interpretability and classification quality. In [14], a framework is proposed in which the graph representation of text is supplemented with symbolic feature selection, while in [15], a hierarchical graph framework is developed that combines linguistic features, domain ontology, multilayer GNN training, attention mechanisms, and dynamic fusion with BERT.

The concept of information ontological modeling proposed in [16] considers information retrieval and clustering methods as stages of the formation of ontological models and the extraction of implicit knowledge.

In [17], the authors investigate the dependency of the topological characteristics of a word graph on the text genre. Texts of different genres are represented as graphs, in which vertices correspond to words, while edges correspond to their neighborhood in bigrams. The parameters of such networks, which are shown to systematically vary between genres, can be used to characterize these genres as well as to identify subgenres. In [18], the ICAN model is used to construct semantic graphs at the level of individual documents, thus allowing word order to be taken into account, which is important for semantic analysis.

2. ONTOLOGICAL BASIS AND DATA CORPUS

2.1. Passports of VAK specialties

In the Russian Federation's scientific personnel certification system, the texts of the VAK scientific specialty passports serve as normative descriptions of subject areas. Each passport is assigned to a specific specialty code to define a set of typical research objects, problem classes, applied methods, and areas of application of the results. These documents have supradepartmental status and serve as the formal basis for classifying dissertations, abstracts, and publications within a specific scientific field.

In terms of content, passports have a stable internal structure, in which, as a rule, the following blocks are distinguished:

- list of main areas and objects of research;
- typical scientific tasks and methods;
- related areas of knowledge.

This structure enables the use of specialty passports not only as regulatory documents but also as textual representations of subject area ontologies. Here, ontology is understood as an explicitly defined system of concepts and relationships that delimits the permissible subject space.

The set of specialty passports can be interpreted as a finite set of ontological objects:

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

where each element s_i corresponds to one scientific specialty, and the text of the passport defines its conceptual content.

The hierarchical structure of codes (aggregated fields, groups, specific specialties) defines a partial order on the set O to capture inclusion and proximity relationships between fields. An ontological space of scientific specialties is introduced in which the level of aggregated fields corresponds to more general classes, while individual codes are used to define specialized subdomains.

Further, a key methodological assumption is adopted. The passport text is viewed as a prototype of the corresponding subject area. Lexemes and set phrases are interpreted as superficial realizations of the concepts and relations of this area, while the thematic coherence of these elements is reflected in the structure of sections. This enables the corpus of scientific specialty passports

[12] A clustering algorithm in machine learning based on the concept of messaging between data points.

[13] 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.

[14] 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.

to be used to construct a supporting semantic-ontological feature space that describes both the passports themselves and the documents being analyzed.

At the same time, passports are not formal ontologies in the strict logical sense and as such cannot be used to define a system of axioms or strict constraints. However, due to their official status, structural stability, and focus on descriptions of objects, tasks, and methods, they can be considered as a practically acceptable ontological basis for the information analysis of scientific texts.

In what follows, this basis will be used to construct a semantic-ontological model of features, as well as to determine the information morphism between documents and the space of scientific specialties.

2.2. Collection and storage of text data

The ontological foundation corpus is formed from the scientific specialty passports approved by the VAK. These specialty passports are collected using specialized software parsers that automatically crawl linked web pages of official resources, download data sheet files and associated documents (in HTML and PDF formats), and capture technical metadata.

For PDF documents, the presence of a text layer is checked. If a text layer is present, extraction is performed directly; otherwise, optical character recognition (OCR) is used, followed by basic cleanup of the results.

At this stage, the text is not yet linguistically normalized, but obvious artifacts (random binary insertions, incorrect encodings) are removed to ensure the correctness of further processing. A detailed normalization algorithm is described below.

For each passport, the following fields are stored, in particular:

- unique record identifier,
- specialty code,
- name,
- uniform resource locator (URL) original source,
- collection time,
- file hash value,
- extracted text after primary cleaning.

The primary passport corpus is generated as JSONL[15] files (one object per line), which can be conveniently used for archiving and data exchange. For long-term storage and efficient access, the data is transferred to the PostgreSQL[16] relational database management system.

For specialty passports, a separate table is created that includes the code and name of the specialty, source metadata, file checksum, and normalized passport text used in constructing features.

For input documents (scientific articles and dissertation abstracts), a corresponding table is created in which, in addition to the text content and document type, the source code of the specialty (for dissertation abstracts), if available, is recorded, since it is used as an expert assessment when assessing the quality of the model.

2.3. Text normalization

Text normalization plays a key role in the proposed model, since it is at this stage that a single feature space is defined, in which scientific texts will subsequently be compared with the passports of scientific specialties.

The processing of text content is deliberately organized as common to all categories of documents in such a way that differences in features reflect the substantive features of the texts, rather than artifacts of format, layout, or source.

The first stage involves technical cleaning of the extracted text. After extracting content from HTML pages or PDF files (using OCR if there is no text layer), incorrect encodings, binary inserts, and fragments of service markup are removed. Line breaks are then converted to regular spaces. For PDF documents, duplicate headers and footers are automatically detected and removed.

Next, service blocks are filtered. Clearly editorial and publishing fragments (References, Keywords, Publisher Information, Author Contact Information, Author Information, etc.) are removed from the texts, as well as technical elements such as URLs, e-mail addresses, and numeric identifiers (ORCID[17], Scopus ID[18], etc.).

The next step is text conversion. All text is converted to lowercase, the “ё” character is replaced with “е”, hyphens are eliminated, and extra spaces are removed. Punctuation and all special characters are removed. Particular attention is paid to hyphenated terms, as a splicing dictionary is used to maintain terminological integrity. According to this dictionary, compound words like object-oriented are converted to underscored (object_oriented) forms. This enables stable compound terms to be considered as single lexical units in subsequent analysis.

Next, numeric expressions are processed. If a token consists solely of digits, it is converted to its Russian verbal form (for example, “2025” is converted to “two thousand twenty-five”), which prevents the dictionary from growing too large due to the large number of unique

[15] JSON Lines is a text format in which each line contains one valid JSON object (JavaScript object notation).

[16] https://www.postgresql.org/. Accessed April 04, 2026.

[17] Open researcher and contributor identifier.

[18] Unique author identifier in the Scopus database. https://www.scopus.com/. Accessed April 04, 2026.

numbers while preserving the semantic information about quantitative characteristics. However, technical standards, indices, and designations (e.g., “5g”, “3d”, “h264”, “ipv6”) are stored as separate tokens and normalized to lowercase, as they are important domain markers.

The linguistic normalization stage involves tokenization, lemmatization, and stopword filtering. The text is broken down into tokens, taking into account previously established rules for punctuation and hyphenation replacement. A normal form (lemma) is determined for each token. Where necessary, common abbreviations and unit designations are standardized, and domain-specific abbreviations are converted to standard forms.

The final result of normalization is a stable text representation as a sequence of lemmas separated by spaces, excluding punctuation, numbers, and function words. If necessary, a list of tokens with positional ranges in the source text is additionally generated, thus enabling the subsequent interpretation of the contribution of individual fragments to the final proximity scores. It is this normalized sequence of lemmas that is used to construct features in the lexical (TF-IDF) and semantic (ICAN) layers of the model.

3. GRAPH SEMANTIC-ONTOLOGICAL MODEL OF FEATURES

3.1. Passport chunks and *n*-grams

The normalized texts of scientific specialty passports (see 2.3) are further considered as sequences of lemmas w:

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

where T_s is the length of passport s in tokens after pre-processing.

Since the specialty passports consist of heterogeneous sections (description of objects, tasks, methods, interdisciplinary connections), then representing the entire text with one vector leads to the fact that weakly related fragments are averaged, which reduces the sensitivity of the model to thematic differences within the passport.

To increase the model’s sensitivity to local areas of semantic concentration, each passport is divided into fragments of a fixed length of L token chunks. The number of chunks for a passport s is defined as follows:

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

Each chunk $c_{s,i}$ inherits the passport metadata (code and name of the specialty), as well as storing the positional range $[a_i, b_i]$ in the source text. This enables the subsequent interpretation of the contribution of individual fragments to the final assessments.

The choice of length L is determined by a trade-off between local sensitivity and stability of the representation.

Chunks that are too short lead to excessive sparsity in the vector space, which increases the impact of markup artifacts and OCR errors on each individual fragment. Chunks that are too long, on the other hand, make local similarity metrics virtually equivalent to their global equivalents, effectively negating the effect of text chunking.

In the conducted experiments, the values of $L \in \{100, 200, 300\}$ were analyzed; the length $L = 200$, which demonstrates a balance between the number of fragments per passport, noise resistance and computational cost, is further used as the default value (Figure).

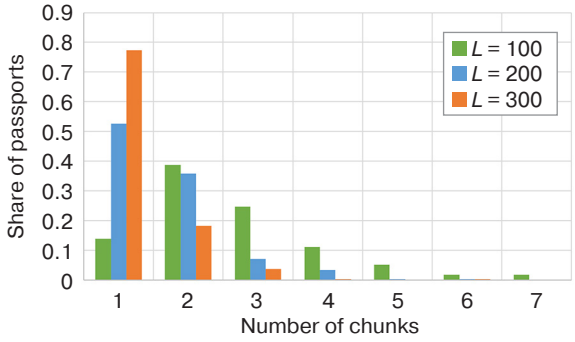

Figure. Distribution of the number of chunks n_s by passports at $L \in \{100, 200, 300\}$

Chunking achieves two goals. Firstly, it localizes specialized vocabulary. Specifically, within a fragment of hundreds of tokens, the share of terms characteristic of a specific field increases, leading to sharper cosine similarities when compared with documents. Secondly, it increases resilience to structural and layout artifacts, in which individual technical inserts or remnants of headers and footers only affect their local vectors and do not dominate the global assessment.

Therefore, the presence of at least two or three fragments in a significant share of passports is important for the local metric, since it increases the probability that at least one fragment will be close to the document in question, while making the aggregate by top N more sensitive to thematic matches.

Next, a lexical dictionary of n-grams is constructed on the set of all passport chunks $D = \{c_1, \ldots, c_N\}$. For each chunk $c \in D$, there is a sequence of lemmas $(w_1, \ldots, w_{T(c)})$.

Two levels are considered:

- set of unigrams $U(c) = \{w_i\}_{i=1}^{T(c)}$,
- set of bigrams $B(c) = \{(w_i, w_{i+1}\}_{i=1}^{T(c)-1}$.

The use of bigrams is essential for capturing stable terminological combinations (neural network, Internet network, communication network, etc.), which clarify the semantics of polysemous lemmas and serve as more reliable markers of subject areas than individual words.

For each term t (both unigrams and bigrams), the number of corpus chunks in which this term occurs at least once is calculated:

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

Next, threshold restrictions are introduced for the minimum and maximum number of occurrences:

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

where d_{MIN} sets the lower limit of the number of chunks in which the term must occur, which allows for the preservation of rare but substantively significant lexemes, while d_{MAX} sets the upper limit of the relative prevalence of the term in the corpus, excluding excessively frequent, uninformative units from the dictionary.

The resulting dictionary V is defined as the set of terms that satisfy the specified conditions:

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

where V′ is the set of all unigrams and bigrams extracted from the corpus of chunks.

Thus, in the semantic-ontological model of features of the specialty passport, a basic space of n-grams V is defined in which TF-IDF representations of both the passports themselves (at the level of chunks and centroids) and the input documents can then be constructed.

The use of chunking and a combined dictionary of unigrams and bigrams allows us to combine local thematic sensitivity with noise resistance and ensure the interpretability of the resulting feature vectors.

3.2. Lexical layer

At the lexical level, each chunk of a specialty passport is considered as a document in a general dictionary of n-grams V, constructed from the passport corpus (see Section 3.1). For each term $t \in \mathrm{V}$ and chunk c, the occurrence frequency $\mathrm{tf}(t, c)$ is calculated, after which a weighted representation is formed using the TF-IDF algorithm.

To account for repeated occurrences of terms in a document, the sublogarithmic term frequency (sublogarithmic TF) is used:

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

The use of sublogarithmic TF allows the contribution of frequently repeated terms within a single fragment to be reduced: since each subsequent occurrence increases their weight with decreasing increment, the influence of local abnormally high frequencies is lowered along with the dependence of the score on the length of the text fragment.

The inverse frequency of the IDF is calculated using the smoothed formula:

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

where N is the total number of chunks in the corpus of passports, $df(t)$ is the number of chunks in which the term t occurs at least once.

Adding one to the numerator and denominator ensures numerical stability at extreme values of $df(t)$, while a shift of one keeps the weights positive even for terms that appear in all documents. Moreover, the $\mathrm{idf}(t)$ function remains monotonically decreasing as $df(t)$ increases, with rare terms receiving higher weights and frequently occurring terms receiving lower weights.

The weight of a term t in a chunk c is given by the product:

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

The vector $\mathbf{w}(c)$, the values of which are given by the quantities $\mathrm{w}_t(c)$, is interpreted as an ontological vector, where each coordinate corresponds to a lexeme or stable phrase associated with an ontology element (a concept, property, or typical relation in the subject area).

Thus, the lexical layer captures the contribution of ontological markers that appear at the level of terms and n-grams.

To ensure comparability of vectors, L_2 normalization is performed:

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

where $\|\mathbf{w}(c)\|_2$ denotes the length of the vector.

Following L_2 normalization, the consideration of the total length and volume of text when comparing documents is significantly reduced; here, the direction of the vector is determined by the relative distribution of term weights.

Since the TF-IDF values are non-negative, the scalar product of two L_2-normalized vectors coincides with the cosine similarity and lies in the range [0; 1], which is convenient for subsequent interpretation of the results.

The normalized TF-IDF vectors of all passport chunks form a sparse matrix

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

where $M = \Sigma_s n_s$ is the total number of chunks for all specialties, $|\mathrm{V}|$ is the dictionary size.

The matrix is stored in the compressed sparse row (CSR) format, which stores the array of nonzero weights, indices of the corresponding terms, and row boundary pointers separately. Using CSR format ensures compact storage, as well as enabling the $v^T\mathbf{X}$ product to be calculated in time proportional to the number of nonzero components of the vector v and nonzero elements of the matrix **X**.

Subsequently, input documents (scientific articles, dissertation abstracts) are represented in the same dictionary V and with the same idf(t) values. Their normalized TF-IDF vectors are compared with chunk vectors and aggregated passport representations, with cosine similarities serving as the basic information-ontological metrics in the proposed model.

3.3. Semantic layer

While the lexical layer captures the coincidences of terms and set phrases, it remains sensitive to variability in wording, synonymy, and abbreviations. To account for contextual relationships between lemmas, a semantic layer is introduced based on the ICAN graph model, in which each text is described as a local term co-occurrence graph [19].

Let the normalized text of a document (a chunk of a passport, article, or abstract) be given by a sequence of lemmas $d = (t_1, \ldots, t_T)$, and the set of unique lemmas of this text be designated as $W = \{u_1, ..., u_{m_d}\}$. Then, a directed weighted graph G_d is constructed for the text on the vertices W, whose structure is given by an adjacency matrix $\mathbf{M} \in [0;1]^{m_d \times m_d}$, initialized to zeros.

The graph is formed by traversing the text with a sliding window of fixed width W (by default, W = 11 tokens). At each position in the window, the central token x is selected, while all other tokens in the window are considered as its contextual neighbors y. The matrix **M** is updated in three stages.

In the first stage, direct connections between the central token and its context (contextual neighbors to the left and right of the central token x) are strengthened.

If the connection $M_{xy} = 0$, then the connection is initialized with a base weight of 0.5, and upon repeated occurrences in the same context relationship, the weight smoothly increases according to formula (12) to ensure monotonous growth at the same time as limiting the value of the weight.

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

In the second stage, second-order indirect connections are taken into account. If token y is already connected to the token k (i.e., $M_{yk} > 0$), then a weakened contribution is added to the pair (x, k):

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

where $A \ll 1$ is the scaling factor.

As a result, the resulting graph reflects not only the co-occurrence of terms in a single sliding window, but also connections through common contextual neighbors, which enables the capture of broader contextual associations.

In the third stage, a damping and thresholding operation is applied to the matrix **M**. All weights are multiplied by a coefficient $\gamma \in (0, 1)$, and elements with a value below the threshold θ are set to zero:

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

where $M_{xy} = 0$ for $M_{xy} \le \theta$, $\gamma = 0.9$ is the attenuation coefficient, $\theta = 0.4$ is the threshold coefficient for removing the connection.

This procedure suppresses random weak connections to form a more stable graph structure that reflects stable contextual connections.

Ultimately, the semantic representation of the text in the ICAN space is defined by a vector of vertex degrees. For each lemma $u_i \in W$, its degree (the sum of the weights of its outgoing and incoming edges) is calculated):

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

Vector $\mathbf{k}(d) = (k_1, ..., k_{m_d})$ reflects the relative importance of lemmas in the contextual structure of the text in which high values correspond to terms that play the role of "nodal points" of the semantic graph.

To integrate with the lexical layer and ontological foundation, the vector $\mathbf{k}(d)$ is projected onto the common basis of the dictionary V constructed from the passport corpus. If $V = \{v_1, \ldots, v_{|V|}\}$, then the semantic vector $\mathbf{s}(d) \in \mathbb{R}^{|V|}$ is determined by the formula:

$$s_j(d) = \begin{cases} k_i, & \text{if lemma } u_i \text{ coincides with } v_j, \\ 0. \end{cases} \quad (16)$$

Next, L_2 normalization is performed:

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

which makes semantic vectors comparable in scale and enables the use of cosine similarity.

4. METRICS OF INFORMATION MORPHISM

Along with the lexical (TF-IDF) layer (previously defined in Section 3.2), the semantic (ICAN) layer (previously defined in Section 3.3), defines two consistent feature spaces, in each of which a document can be compared with scientific specialty datasets. In both cases, two groups of metrics are used:

- local, measuring the proximity of the document to individual fragments of the passport;

- global, characterizing the proximity of the document to the centroid of the passport, obtained by averaging over all its fragments.

These metrics are introduced in a standardized form and then applied separately in each layer.

Let a normalized feature vector $\mathbf{v}(d)$ be constructed for the input document d, and let a normalized vector $\mathbf{u}_{s,i}$, where $i = 1, \ldots, n_s$, be constructed for each chunk $c_{s,i}$ of the passport s. All vectors are normalized and non-negative.

Then the local metric is defined as the cosine similarity of the document with a separate passport chunk vector according to the formula:

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

For each passport s, this defines a set of local assessments $\text{CosSimilarity}(d, c_{s,i})_{i=1}^{n_s}$, which reflect how closely the document is comparable in content to individual fragments of the passport text.

Based on these estimates, an aggregated local metric MaxSim_k, is introduced, taking into account only a few of the largest matches

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

and by $r_{s,1} \geq \ldots \geq r_{s,n_s}$ are the values ordered in descending order. Then the local estimate for s is given as:

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

Thus, MaxSim_k reflects the presence of multiple fragments in a passport that are as close as possible to the document's content. The choice of the parameter k provides a compromise between sensitivity to narrow matches and robustness to noise, while in the experimental section, a value of three ($k = 3$) is used.

The global metric is formed using the passport centroid in the layer under consideration. For passport s, the average vector is determined by the formula:

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

After this, normalization is performed; here, the global measure of closeness of document d to the passport s is given by the cosine similarity with the centroid according to the formula:

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

As a result, in each layer, documents and specialty passports are linked by a pair of complementary metrics: $\text{MaxSim}_k(d, s)$ is a local metric sensitive to the most relevant fragments of the passport; $\text{CentroidCos}(d, s)$ is a global metric reflecting compliance with the general topic.

4.1. Hybrid model and integral metric

Local and global metrics defined for the lexical and semantic layers provide consistent estimates of the closeness of a document d to a passport s.

These metrics are then combined into a single scalar value within the corresponding layer, after which the results of the lexical and semantic layers are aggregated into a final hybrid metric.

Within each layer, the local and global metrics are combined using the formula:

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

where coefficient α sets the balance between $\text{CentroidCos}(d, s)$, indicating the proximity to the passport as a whole, and $\text{MaxSim}(d, s)$, which highlights the closest fragments of the passport.

In the TF-IDF lexical layer, increasing α enhances the contribution of exact terminological matches. This is due to the fact that, at $\alpha > 0.5$, the influence of local maxima of MaxSim, which arise in those passports that contain fragments with a high concentration of matching n-grams, increases. The value of $\alpha = 0.6$ was chosen empirically based on the validation results as providing the best quality of classification and recommendation ranking.

The resulting $S(d, s)$ values for TF-IDF and ICAN are interpreted as two independent but consistent assessments of document compliance with the passport. The first interpretation consists in terms of precise lexical markers, while the second is considered in terms of contextual associations. A hybrid metric is introduced to form the final scalar score:

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

where the parameter λ regulates the contribution of the semantic layer relative to the lexical layer.

At λ approaches 1, ICAN graph representations dominate, increasing robustness to paraphrases and abbreviations. In the experiments conducted, the values of $\alpha_{TF} \approx 0.6$, $\alpha_{ICAN} \approx 0.5$, $\lambda \approx 0.5$ were empirically selected to provide a balance between accuracy and robustness on abstract and article corpora.

From a computational point of view, all components of the metric $S(d, s)$ are obtained from two vector-matrix operations of the form $\mathbf{r} = \mathbf{X} \times \mathbf{v}(d)$, where $\mathbf{X}$ is a matrix of chunk vectors, while $\mathbf{r}$ is a vector of cosine similarities with all chunks of all passports.

Formally, from a conceptual point of view, the value of S(d, s) can be interpreted as the intensity of information morphism from document d to ontological entity s: the higher the S(d, s), the more consistently the content of the text reproduces the lexical-semantic profile of the corresponding specialty.

4.2. Evaluation of probabilistic morphisms

The integral metric S(d, s) defines a scalar conformity score for each document d and specialty passport $s \in$ O. In order to transition from a set of such scores to a formal information morphism, it is necessary to construct a probability distribution over the ontological set of passports O.

Let the values $S(d,s)_{s \in O}$ be calculated for a fixed document d. Then a probability distribution is introduced based on them using the softmax scheme with temperature $\tau > 0$:

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

where $M = \max_{q \in O} S(d,q)$ is used for numerical stabilization (log-sum-exp normalization). Subtracting M does not affect the relative probabilities, but prevents overflow when carrying out exponential transformations.

Parameter τ determines the degree of concentration of the probability distribution:

- when $\tau \ll 1$, distribution becomes more concentrated, since even small differences in the values of S(d, s) lead to a pronounced dominance of one or more passports;
- when $\tau \approx 1$, standard temperature scaling is applied, in which softmax preserves the typical normalization behavior of estimates;
- when $\tau > 1$, the distribution becomes smoother; this is convenient when analyzing interdisciplinary and borderline texts, which are characterized by the presence of several estimates comparable in magnitude.

The resulting distribution $P(s \mid d, \tau)$ defines a mapping:

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

where D is a set of documents, and Δ(O) is a simplex of probability measures on a set of specialty passports.

Here μ is interpreted as an information morphism from the text space to the ontological space O, where each document is matched with the distribution of its thematic contribution across normatively defined areas.

The point solution of the classification problem corresponds to the passport with the highest probability:

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

5. EXPERIMENTAL EVALUATION

5.1. Results of model validation based on applicants' abstracts

The model was validated using a corpus of 124 abstracts submitted by applicants. The corpus covers sixteen dissertation councils from five organizations: MIREA – Russian Technological University (RTU MIREA)[19], V.F. Utkin Ryazan State Radio Engineering University[20], Federal Research Center for Informatics and Management of the Russian Academy of Sciences[21], V.A. Trapeznikov Institute of Control Sciences of the Russian Academy of Sciences[22], and Patrice Lumumba Peoples' Friendship University of Russia (RUDN)[23].

For each council, abstracts on codes falling within its area of competence were taken into account, for example, 2.3.2, 2.3.5, 2.3.8 for council D24.2.326.09. The official code of the specialty specified in the dissertation council data was taken as the reference mark for each document.

For three configurations of the lexical TF-IDF, semantic ICAN, and hybrid models, two metrics were evaluated. The first metric is top 1 accuracy, i.e., the share of abstracts for which the passport with the maximum value S(d, s) coincides with the reference code. The second metric is top 3 accuracy, i.e., the share of abstracts for which the reference code is among the top three passports with the highest S(d, s) values.

The aggregation of values across the entire corpus demonstrates the consistent advantage of the hybrid model. The average top 1 accuracy across all councils and documents is approximately 0.58 for TF-IDF, approximately 0.57 for ICAN, and approximately 0.69 for the hybrid configuration. For top 3 accuracy, the picture is even more pronounced. The TF-IDF model gives an average of about 0.74, while ICAN gives around 0.81, and the hybrid model reaches about 0.90. That is to say, in nine out of ten cases, the correct passport is among the top three closest in value to S(d, s).

A review of the results for individual organizations shows a similar picture. In the RTU MIREA samples (Table 1), where councils D24.2.326.09, D24.2.326.03, D24.2.326.08, and D24.2.326.10 were analyzed, the average top 1 accuracy for abstracts is

[19] https://www.mirea.ru/. Accessed April 04, 2026. (In Russ.).
[20] https://rsreu.ru/. Accessed April 04, 2026. (In Russ.).
[21] https://www.frccsc.ru/. Accessed April 04, 2026. (In Russ.).
[22] https://www.ipu.ru/. Accessed April 04, 2026. (In Russ.).
[23] https://www.rudn.ru/. Accessed April 04, 2026. (In Russ.).

approximately 0.54 for TF-IDF and about 0.63 for ICAN. The hybrid model increases this indicator to 0.75.

According to the top 3 metric, the hybrid configuration reaches about 0.96, i.e., in almost all cases, the correct specialty code is among the three most probable.

At V.F. Utkin Ryazan State Radio Engineering University (Table 2), the lexical model for councils D24.2.375.01, D24.2.375.02, D24.2.375.03, and D99.2.113.02 demonstrates an average top 1 accuracy of 0.68, while the corresponding value for ICAN is about 0.61.

At the same time, the semantic layer provides higher accuracy in the top 3, i.e., approximately 0.89 compared to 0.78 for TF-IDF.

The hybrid model combines the advantages of both configurations to achieve an average top 1 accuracy of around 0.75 and top 3 accuracy of around 0.93.

At the Federal Research Center for Informatics and Management of the Russian Academy of Sciences (Table 3), the average top 1 accuracy for councils D24.1.224.04, D24.1.224.03, and D24.1.224.02 is about 0.60 for TF-IDF and about 0.50 for ICAN. The hybrid model increases this indicator to 0.70.

At the same time, according to the top 3 TF-IDF and ICAN metrics, similar values of around 0.83 are achieved, while the hybrid configuration reaches a level of around 0.90.

Table 1. Results of the model for authors' abstracts at RTU MIREA

Council D24.2.326.09			Council D24.2.326.03		
Scientific specialties 2.3.2, 2.3.5, 2.3.8			Scientific specialties 1.4.7, 1.4.10		
Method	Accuracy according to top 1	Accuracy according to top 3	Method	Accuracy according to top 1	Accuracy according to top 3
TF-IDF	0.6	0.7	TF-IDF	0.5	0.63
ICAN	0.7	0.7	ICAN	0.5	0.75
Hybrid	0.8	1	Hybrid	0.63	0.88
Council D24.2.326.08			Council D24.2.326.10		
Scientific specialties 1.2.2, 2.3.1			Scientific specialty 5.2.3		
Method	Accuracy according to top 1	Accuracy according to top 3	Method	Accuracy according to top 1	Accuracy according to top 3
TF-IDF	0.33	0.67	TF-IDF	0.67	0.67
ICAN	0.67	0.67	ICAN	0.67	1
Hybrid	0.67	1	Hybrid	1	1

Table 2. Results of the model for authors' abstracts in V.F. Utkin Ryazan State Radio Engineering University

Council D24.2.375.01			Council D24.2.375.02		
Scientific specialties 2.3.1, 2.3.5			Scientific specialties 1.3.2, 1.3.11, 2.2.1		
Method	Accuracy according to top 1	Accuracy according to top 3	Method	Accuracy according to top 1	Accuracy according to top 3
TF-IDF	0.67	0.67	TF-IDF	0.67	0.83
ICAN	0.67	1	ICAN	0.5	0.83
Hybrid	0.67	0.67	Hybrid	0.5	0.83
Council D24.2.375.03			Council D99.2.113.02		
Scientific specialties 2.2.11, 2.2.12, 2.2.13			Scientific specialty 2.3.8		
Method	Accuracy according to top 1	Accuracy according to top 3	Method	Accuracy according to top 1	Accuracy according to top 3
TF-IDF	0.69	0.81	TF-IDF	0.67	0.67
ICAN	0.63	0.94	ICAN	0.67	0.67
Hybrid	0.81	1	Hybrid	1	1

Table 3. Results of the model for authors' abstracts at the Federal Research Center for Informatics and Management of the Russian Academy of Sciences

Council D24.1.224.04			Council D24.1.224.03			Council D24.1.224.02		
Scientific specialties 2.3.2, 2.3.5, 2.3.6			Scientific specialties 1.2.1, 1.2.3, 2.3.8			Scientific specialties 1.1.2, 1.1.6, 1.1.9		
Method	Accuracy according to top 1	Accuracy according to top 3	Method	Accuracy according to top 1	Accuracy according to top 3	Method	Accuracy according to top 1	Accuracy according to top 3
TF-IDF	0.67	1	TF-IDF	0.56	0.78	TF-IDF	0.67	0.89
ICAN	0.67	1	ICAN	0.39	0.83	ICAN	0.67	0.78
Hybrid	1	1	Hybrid	0.61	0.89	Hybrid	0.78	0.89

At the V.A. Trapeznikov Institute of Control Sciences of the Russian Academy of Sciences (Table 4), where the corpus includes five abstracts from council D24.1.107.02, the results should be interpreted with caution due to the small sample size. In this group, the ICAN semantic model gives the best top 1 accuracy of about 0.60 compared to 0.40 for TF-IDF.

While the hybrid configuration maintains a top 1 level of around 0.60, it does not surpass the semantic branch in terms of top 3.

At RUDN (Table 5), the average top 1 TF-IDF and ICAN accuracy for four councils PDS 0300.004, PDS 2028.001, PDS 0900.006, and PDS 0200.002 is close, amounting to approximately 0.54.

The hybrid model increases this indicator to 0.62. For the top 3 lexical and semantic configurations, the values are around 0.70 and 0.78, while the hybrid model reaches about 0.89. Significant gains are observed on individual councils.

Table 4. Results of the model for authors' abstracts at the V.A. Trapeznikov Institute of Control Sciences of the Russian Academy of Sciences

Council D24.1.107.02		
Scientific specialties 2.3.1, 2.3.4		
Method	Accuracy according to top 1	Accuracy according to top 3
TF-IDF	0.4	0.6
ICAN	0.6	0.8
Hybrid	0.6	0.6

For PDS 2028.001, the hybrid model improves the top 3 accuracy from 0.56 for TF-IDF and 0.78 for ICAN to unity. For PDS 0200.002, the hybrid configuration also increases the top 3 accuracy to unity while maintaining high top 1 accuracy.

Table 5. Results of the model for authors' abstracts at RUDN University

Council PDS 0300.004			Council PDS 2028.001		
Scientific specialties 3.1.18, 3.1.20, 3.3.6			Scientific specialties 5.8.1, 5.8.7		
Method	Accuracy according to top 1	Accuracy according to top 3	Method	Accuracy according to top 1	Accuracy according to top 3
TF-IDF	0.56	0.75	TF-IDF	0.44	0.56
ICAN	0.44	0.81	ICAN	0.44	0.78
Hybrid	0.56	0.81	Hybrid	0.56	1
Council PDS 0900.006			Council PDS 0200.002		
Scientific specialty 5.1.4			Scientific specialties 1.4.1, 1.4.3, 1.4.4		
Method	Accuracy according to top 1	Accuracy according to top 3	Method	Accuracy according to top 1	Accuracy according to top 3
TF-IDF	0.4	0.6	TF-IDF	0.71	0.86
ICAN	0.8	0.8	ICAN	0.71	0.71
Hybrid	0.8	0.8	Hybrid	0.71	1

Table 6. Results of the evaluation of scientific articles by broad scientific fields

Method	Accuracy for each broad scientific field				
	1. Natural sciences	2. Engineering sciences	3. Medical sciences	4. Agricultural sciences	5. Social sciences and humanities
TF-IDF	0.85	0.87	0.85	0.86	0.83
ICAN	0.79	0.72	0.74	0.9	0.91
Hybrid	0.9	0.88	0.93	0.92	0.94

Thus, validation on the corpus of applicants' abstracts demonstrates the stable and interpretable quality of specialty code recovery using the proposed semantic-ontological model in a hybrid configuration.

In most cases, the correct passport ends up in a narrow set of the most likely candidates, making the model suitable both for automated support in selecting a specialty code and for analyzing alternative thematically related areas.

5.2. Results of model validation in scientific articles published in the VAK list

The model was validated using scientific articles from a corpus of publications from journals included in the current VAK list. Since these journals are known to represent broad scientific fields and groups of scientific specialties, it was possible to conduct an assessment both at the aggregate level of fields and at the level of passport groups.

The first experiment analyzed the accuracy of assigning an article to a broad field of science. After determining the reference field based on the journal's profile in the VAK list, the predicted field was calculated based on the passport with the maximum value of S(d, s). The results show differences in the behavior of lexical, semantic, and hybrid configurations (Table 6).

The TF-IDF method demonstrates the greatest stability in natural and engineering sciences, where the top 1 accuracy is 0.85 and 0.87, respectively. In medical and agricultural sciences, the accuracy of TF-IDF is 0.85 and 0.86, respectively, while in social sciences and humanities, it is 0.83. This is consistent with the fact that in natural sciences and technical fields, terminology is more standardized and closer to the wording of passports.

The ICAN semantic model performs significantly better in cases where the language of publications is more variable. In the social sciences and humanities, its accuracy reaches 0.91, compared to 0.83 for TF-IDF. In agricultural sciences, ICAN scores 0.90 compared to 0.86 for TF-IDF. The slight inferiority of ICAN to the lexical model in natural, engineering, and medical sciences reflects the dependence of the purely semantic layer on the quality of local graphs in terminology-rich but well-standardized fields.

The hybrid configuration combines the strengths of both approaches. Across all five broad areas, it yields the highest top 1 accuracy values. For natural sciences, the accuracy is 0.90; for engineering, 0.88; for medicine, 0.93; for agriculture, 0.92; and for social sciences and humanities, 0.94. Thus, when moving to a higher level of aggregation, the hybrid model almost always corrects the errors of each of the single configurations to ensure the most stable behavior.

The second experiment evaluated the accuracy of restoring groups of scientific specialties. For each article, a vector S(d, s) was formed for all passports, after which the predicted group was determined by the passport with the maximum value of S(d, s). The reference group was set according to the declared specialization of the journal. Here, the top 1 and top 3 metrics were analyzed (Table 7).

Table 7. Results of the evaluation of scientific articles by groups of scientific specialties

Method	Accuracy according to top 1	Accuracy according to top 3
TF-IDF	0.9	0.94
ICAN	0.87	0.97
Hybrid	0.91	0.96

At the specialty group level, the TF-IDF lexical model achieves a top 1 accuracy of 0.90 and a top 3 accuracy of 0.94. The ICAN semantic model gives a slightly lower top 1 accuracy of 0.87, but a higher top 3 accuracy of 0.97. This means that ICAN is slightly more likely to make a mistake in selecting the single closest passport, but almost always includes the correct group among the three most likely options.

The hybrid model retains the best top 1 value of 0.91, while providing high top 3 accuracy of 0.96. While not outperforming ICAN in terms of top 3, it provides a more balanced ratio between the accuracy of the first choice and the completeness of the top three groups. In the context of recommendation ranking tasks based on codes, this configuration is the most practical, since the main code can be reliably suggested to simultaneously form a content-relevant list of alternative specialties.

Overall, the results for articles from journals included in the VAK list confirm the conclusions made based on the abstracts. The lexical model works best in areas with a rigidly defined nomenclature, while the semantic model is particularly useful in the humanities and related fields, and the hybrid configuration provides the most stable quality at all levels of aggregation to provide interpretable probability profiles by scientific specialty group.

5.3. Comparison of a graph-based semantic-ontological model with large language models

To evaluate the proposed model, we compared it with a number of large language models applied in classification mode without retraining. In all cases, the same task was established: to restore the specialty code or the closest passport based on the text of the document. Top 1 accuracy and top 3 accuracy were calculated on corpora of abstracts and scientific articles (Table 8).

According to the abstracts, the hybrid graph semantic-ontological model gives a top 1 accuracy of 0.69 and a top 3 accuracy of 0.90. The lexical TF-IDF model and semantic ICAN are inferior to it in both metrics (0.58 for TF-IDF and 0.74 for ICAN, 0.57 and 0.81, respectively). Among large language models, the ChatGPT 5.2 Thinking[24] configuration achieves the best values, with a top 1 accuracy of 0.79 and a top 3 accuracy of 0.84. The ChatGPT 4o[25] model performs at 0.71 and 0.73, DeepSeek[26] at 0.61 and 0.70, LLaMA[27] at 0.57 and 0.63, while Alice AI (YandexGPT)[28] lags significantly behind all options, scoring 0.46 and 0.52. Thus, while the large language model ChatGPT 5.2 Thinking outperforms the hybrid scheme in terms of first choice accuracy according to the abstracts, the graph semantic-ontological model provides higher completeness for the three closest specialties to form a fuzzy but substantively stable profile of similarities based on passports.

Table 8. Results of comparing a graph-based semantic ontology model with large language models

Authors' abstracts			Articles		
Method	Accuracy according to top 1	Accuracy according to top 3	Method	Accuracy according to top 1	Accuracy according to top 3
TF-IDF	0.58	0.74	TF-IDF	0.85	0.86
ICAN	0.57	0.81	ICAN	0.81	0.84
Hybrid	0.69	0.9	Hybrid	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
Alice AI (YandexGPT)	0.46	0.52	Alice 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

[24] https://openai.com/ru-RU/index/introducing-gpt-5-2/. Accessed April 04, 2026. (In Russ.).
[25] https://openai.com/ru-RU/index/hello-gpt-4o/. Accessed April 04, 2026. (In Russ.).
[26] https://www.deepseek.com/. Accessed April 04, 2026.
[27] https://www.llama.com/. Accessed April 04, 2026.
[28] https://alice.yandex.ru/. Accessed April 04, 2026. (In Russ.).

According to scientific articles from journals listed in the VAK catalog, the picture is different. The hybrid model achieves a top 1 accuracy of 0.91 and a top 3 accuracy of 0.93. The lexical TF-IDF model gives an accuracy of 0.85 and 0.86, while the corresponding ICAN scores are 0.81 and 0.84, respectively. Among large language models, ChatGPT 4o shows an accuracy of 0.80 and 0.82, while ChatGPT 5.2 Thinking gives 0.82 and 0.97, DeepSeek shows 0.79 and 0.84, LLaMA shows 0.62 and 0.67, and Alice AI shows 0.56 and 0.69. According to the top 1 metric, the hybrid model outperforms all language models by nine to fifteen percentage points. According to the top 3 metric, ChatGPT 5.2 Thinking shows the best result (0.97), while the hybrid model gives a slightly lower value of 0.93, but remaining significantly higher than the other configurations. This means that when classifying articles, the proposed graph semantic-ontological scheme better captures the main passport, while large language models more often include the correct specialty code in a wide set of closest candidates.

The differences in behavior are consistent with the nature of the approaches being compared. The graph-based semantic-ontological model, which is tightly bound to the texts of specialty passports, uses them as an ontological basis to factor solutions into interpretable TF-IDF and ICAN components.

This effect is particularly noticeable in articles from journals included in the VAK list, where the wording is closer to the normative language of passports. Large language models, which rely on generalized representations of scientific disciplines, often take into account context that is not reflected in passport texts, Therefore, in abstracts that contain detailed reviews, references to related fields, and less formalized presentation, the best ChatGPT configurations demonstrate higher accuracy in the first choice, but lose some of their interpretability and controllability.

The comparison shows that the proposed graph semantic-ontological model is comparable to modern language models in terms of classification quality, surpassing them in a number of metrics in scenarios where strict consistency with passport texts is important, while remaining significantly cheaper in terms of computational costs, as well as more transparent in terms of the structure of the decisions made.

Large language models should be viewed as an additional tool that complements rather than replacing ontology-oriented schemas. This is especially relevant in expert support tasks that require both quantitative scoring and explicit linking of results to normative descriptions of scientific disciplines.

6. DISCUSSION OF RESULTS AND CONCLUSION

The graph-based semantic-ontological model of scientific text analysis presented in this work is based on the VAK specialty passports, which are used both as textual descriptions of subject areas and as a source of ontological feature space. After forming a unified dictionary of lemmas and stable phrases based on the texts of the passports, the passports themselves, as well as the dissertation abstracts and scientific articles, can be described in this database.

The similarity of a document to the code of a scientific specialty is interpreted through the consistency of the text content with the normative description of the field.

At the same time, the semantic-ontological space has two coordinated layers. The lexical layer is built on the basis of TF-IDF vectors, which record the use of terms and terminological combinations selected from the passport chunk corpus. The semantic layer, which is based on the ICAN model, describes the text through a graph of co-occurring lemmas and the degree of its vertices. This supports paraphrasing, synonymy, the use of abbreviations, and terms scattered throughout the text, while maintaining a link to the same ontological basis. Thus, the hybrid metric based on lexical and semantic approaches combines local and global metrics in each layer.

Experimental results on abstracts show that the hybrid model systematically outperforms individual lexical and semantic approaches in terms of accuracy. The TF-IDF-based lexical model, which provides high accuracy where terminology is standardized, closely matches the wording of abstracts, while the semantic model is particularly useful in fields involving a freer scientific style, including active use of abbreviations and variable descriptions of the research subject.

Aggregated indicators for broad areas of knowledge confirm the observed pattern. For social sciences and humanities, the semantic layer plays a critical role in improving ranking quality, since thematic boundaries are often expressed through complex formulations and contextual markers. For a significant portion of technical and natural science specialties, the lexical layer already provides a high baseline level of quality, while the semantic layer refines the document profile in cases of similar codes and borderline topics. Thus, the model demonstrates behavior that can be explained in terms of content and is consistent with the peculiarities of the terminological structure of different scientific fields.

Further development of the model can take several directions. One of these involves the formation of lexemes and terminological combinations of passports with external ontologies, knowledge bases, and specialized terminological resources. This will permit a movement

from a textual ontological basis to a more formalized multi-domain ontology. In addition, it is possible to add contextual embeddings and neural network models as an additional layer on top of the existing TF-IDF and ICAN schemas while maintaining interpretability through decomposition on an ontological basis.

A separate task is the optimization of parameters on validation samples and analysis of their dependence on the field of knowledge and document type.

In conclusion, it can be confirmed that the graph-based semantic-ontological model based on the texts of the VAK specialty passports demonstrates the possibility of transforming the normative corpus into a working ontological space for quantitative analysis of scientific texts. The results obtained for abstracts and scientific articles show that this approach can serve as a basis both in the provision of automated support for expert decision-making, as well as for monitoring the structure of scientific knowledge within a given normative ontology.

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.

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. 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.

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

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

Translated from Russian into English by L. Bychkova
Edited for English language and spelling by Thomas A. Beavitt